Contribution à l'analyse et à l'amélioration de certaines méthodes pour l'inférence statistique par vraisemblance pénalisée : Méthodes Bregman-proximales et LASSO

Stéphane Chrétien

Mémoire

déposé en vue de l'obtention de

l'habilitation à diriger des recherches

par

Stéphane Chrétien

Contribution à l'analyse et à l'amélioration de certaines méthodes pour l'inférence statistique par vraisemblance pénalisée: Méthodes Bregman-proximales et LASSO.

et soutenu le 26 novembre 2014 devant le jury composé de

- Pascal Bondon,
- Charles Bouveyron,
- Djalil Chafaï,
- Clément Dombry,
- Florence Forbes,
- Yves Granvalet,

les rapporteurs étant

- Charles Bouveyron,
- Florence Forbes et
- Yves Granvalet.

Contents

Liste des articles et notes

[A] S. Chrétien and P. Bondon, *Cyclic projection methods on a class of nonconvex sets*. Numer. Funct. Anal. Optim. 17 (1996), no. 1-2, 37–56.

[B] S. Chrétien and A.O. Hero, *Kullback proximal algorithms for maximum-likelihood estimation. Information-theoretic imaging.*, IEEE Trans. Inform. Theory 46 (2000), no. 5, 1800–1810..

[C] G. Celeux, S. Chrétien, F. Forbes and A. Mkhadri, *A component-wise EM algorithm for mixtures.*, J. Comput. Graph. Statist. 10 (2001), no. 4, 697–712.

[D] C. Biernacki and S. Chrétien, *Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM.*, Statist. Probab. Lett. 61 (2003), no. 4, 373–382.

[E] S. Chrétien and A.O. Hero, *On EM algorithms and their proximal generalizations.*, ESAIM: PS 12 (2008) 308–326.

[F] S. Chrétien and F. Corset, *Using eigenvalue optimization for binary least squares estimation problems.*, Signal Processing 89 (2009) no. 11, 2079–2091.

[G] S. Chrétien and D. Pleydell, *Mixtures of GAMs for habitat suitability analysis with overdispersed presence / absence data.*, Computational Statistics and Data Analysis 54 (2010), no. 5, 1405–1418.

[H] S. Chrétien, *An alternating ℓ_1 relaxation for compressed sensing*, IEEE Signal Processing Letters 17 (2010) no.2, 181–184.

[I] S. Chrétien, A.O. Hero and H. Perdry, *EM type algorithms for likelihood optimization with non-differentiable penalties*, Ann. Inst. Stat. Math. 64 (2012), no. 4, 791-809

[J] S. Chrétien, *Estimation of Gaussian mixtures in small sample studies using l_1 penalization*, in révision.

[K] S. Chrétien and S. Darses, *Invertibility of random submatrices via tail decoupling and a Matrix Chernoff Inequality* Statistics and Probability Letters, 82 (2012), no. 7, 1479-1487.

[L] S. Chrétien and S. Darses, *Sparse recovery with unknown variance: a LASSO-type approach*, IEEE Trans. Information Theory, 60 (2014), no.7, 3970–3988.

[M] S. Chrétien and J.-P. Ortega, *Multivariate GARCH estimation via a Bregman-proximal trust-region method*, Computational Statistics and Data Analysis, 76 (2014), 210–236.

[N] S. Chrétien, *On prediction with the LASSO when the design is not incoherent*, soumis.

[O] S. Chrétien, *A mixture model for design matrices in the LASSO: weakening incoherence conditions*, en révision.

[P] S. Chretien, J.-M. Nicod, L. Philippe, V. Rehn-Sonigo and L. Toch, *Using a Sparsity Promoting Method in Linear Programming Approximations to Schedule Parallel Jobs*, Concurrency and Computing, à paraître.

[Q] S. Chrétien and S. Darses, *Perturbation bounds on the extremal singular values of a matrix after appending a column, soumis*.

[R] S. Chrétien and S. Darses, *On the spacings between the successive zeros of the Laguerre polynomials*, (with S. Darses), Proceedings of the AMS, à paraître.

Résumé

1 Résumé de la thématique de la thèse d'université

Ma thèse de doctorat intitulée *Méthodes de projections successives pour l'optimisation ensembliste non-convexe* et élaborée sous la direction de Odile Macchi, DR CNRS au sein du Laboratoire des Signaux et Systèmes de l'Université Paris XI - Orsay, est consacrée à l'étude de la convergence de méthodes de projections pour trouver un point dans l'intersection d'ensemble faiblement fermés dans un espace de Hilbert sans hypothèse de convexité.

Historique du problème

Les méthodes de projections successives ont été employées depuis les années 30 pour résoudre itérativement les grands systèmes linéaires et semblent remonter à Kaczmarz (Angenäherte Auflösung von Systemen linearer Gleichungen. Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles. Série A, Sciences Mathématiques, vol. 35, pp. 355–357, 1937). Pour un système

$$Ax = b, \tag{1.1}$$

avec $A \in \mathbb{R}^{m \times n}$, l'idée consiste simplement à projeter successivement sur les ensembles

$$S_j \quad = \quad \left\{ x \mid a_j^t x = b_j \right\} \tag{1.2}$$

$j = 1, \ldots, m$. Plus généralement, lorsqu'on dispose d'une famille de m ensembles S_j faiblement fermés dans un espace de Hilbert $\mathbb{H}$, et si on dénote par P_{S_j} la projection sur S_j, en notant qu'elle est toujours bien définie sur des ensembles faiblement fermés, un algorithme de projections successives aura la forme

$$x^{(l+1)} \quad = \quad P_{S_\sigma^{(l)}(j)} \left(x^{(l)} \right) \tag{1.3}$$

où $\sigma^{(l)} : \{1, \ldots, m\} \mapsto \{1, \ldots, m\}$ est une fonction de choix permettant de selectionner sur quelle contrainte on va projeter à l'itération l. On peut démontrer facilement que la suite $(x^{(l)})_{l \in \mathbb{N}}$ a un point d'accumulation dans $\cap_{j=1}^m S_j$. Le lecteur pourra trouver plus de résultats dans (L.G. Gurin, B.T. Polyak, and E.V. Raik. The method of projections for finding the common point of convex sets. U.S.S.R. Computational Mathematics and Mathematical Physics, 7:1–24, 1967). Par contre, en général, ce point d'accumulation n'est pas une projection du point $x^{(0)}$ sur $\cap_{j=1}^m S_j$. Le cas où les ensembles S_j sont affines et $m = 2$ est très particulier et il est connu depuis Von Neumann (J. von Neumann, On rings of operators. Reduction theory, Ann. of Math. 50 (1949) 401–485) que la suite converge vers la projection de $x^{(0)}$ sur $\cap_{j=1}^m S_j$. Dans le cas plus général des ensembles convexes, une modification due à Dykstra (Boyle, J. P.; Dykstra, R. L. (1986). "A method for finding projections onto the

intersection of convex sets in Hilbert spaces". Lecture Notes in Statistics 37: 28–47) est connue pour converger vers la projection de $x^{(0)}$ sur $\cap_{j=1}^{m} S_j$. Une autre méthode, due à Pierra permet de projeter en utilisant des projections successives sur deux ensembles (G. Pierra, Decomposition through formalization in a product space, Mathematical Programming, 28 (1984) 95–115. Le problème de traiter le cas où les ensembles S_j ne sont pas supposés convexes était relativement ouvert avant d'entreprendre ce travail. Il était motivé principalement par le cas où $\mathbb{H}$ est l'ensemble des matrices dans $\mathbb{R}^{m\times n}$ avec le produit scalaire $\langle A, B\rangle =$ trace(AB^t) et l'un des S_j étant l'ensemble des matrices d'un rang inférieur à un rang donné. En particulier, le problème très important en modélisation et pratique des séries temporelles, consistant à approcher un signal par une somme d'exponentielles complexes, peut se formuler sous cette forme lorsqu'on impose que les matrices concernées soient de plus de type Hankel.

Contributions

Dans une première partie du travail de thèse, je me suis concentré sur le problème dit de faisabilité, c'est à dire simplement de construire une suite $(x^{(l)})_{l\in\mathbb{N}}$ ayant au moins un point d'accumulation dans $\cap_{j=1}^{m} S_j$. J'ai adopté l'approche consistant à définir une classe un peu plus générale que celle des ensembles convexes: les ensembles développables en convexes, c'est à dire consistant en une union dénombrable d'ensembles convexes fermés. Avec cette contrainte sur les ensembles S_j, $j = 1, \ldots, m$, on peut réussir à étendre les résultats de (L.G. Gurin, B.T. Polyak, and E.V. Raik. The method of projections for finding the common point of convex sets. U.S.S.R. Computational Mathematics and Mathematical Physics, 7:1–24, 1967) et garantir ainsi la convergence d'au moins une sous-suite de la suite engendrée par les projections successives pour un schema de sélection des ensembles raisonnable comme par exemple le défilement modulo m. Cette partie a donné lieu en partie à la publication [**alpha**]. La thèse contient deux parties qui particularisent ces résultats à des ensembles plus contraints pour lesquels on obtient une convergence vers un point de l'intersection. Ces résultats n'ont pas fait l'objet de publication autre que dans le manuscrit de thèse.

Dans une deuxième partie, je me suis intéressé à la question de projeter un point $x^{(0)}$ sur l'intersection $\cap_{j=1}^{m} S_j$. J'ai pour cela modifié la méthode de Pierra citée plus haut. J'ai démontré avec Pascal Bondon que sous des conditions géométriquement raisonnables, la modification proposée avait un point d'accumulation qui satisfait bie la propriété recherchée, c'est à dire d'être une projection de $x^{(0)}$ sur l'intersection $\cap_{j=1}^{m} S_j$. Cette partie n'a pas été publiée à l'époque, mais elle a récemment subi des transformations qui permettent de relaxer des hypothèses compliquées à vérifier en pratique. J'ai pu constater combien le recul de quelques années sur un travail peut permettre d'en voir les faiblesses et d'en apprécier la fraicheur. J'ai réalisé dans ce cas précis qu'un projet peut prendre des années à murir et que les éléments manquants pour terminer un travail peuvent parfois surgir de manière inattendue des années plus tard. L'article remanié correspondant [**A18**] sera soumis prochainement pour publication.

2 Résumé des recherches réalisées au cours de la période post-doctorale

Les recherches effectuées au cours des années ayant succédé à l'obtention de la thèse peuvent se diviser selon deux thématiques distinctes. Une première a consisté en une étude approfondie des propriétés de convergence de l'algorithme EM très couramment utilisé en inférence statistique. La deuxième partie des travaux effectués dans cette période s'est concentrée sur

les propriétés de l'approche LASSO (Least Absolute Shrinkage and Selection Operator) très couramment utilisé pour la sélection de variables dans les modèles de régression. Toutes les citations font référence à la liste des publications se situant à la page 6.

3 Algorithme EM pour la maximisation de vraisemblance pénalisée

Les travaux que nous allons présenter ont fait l'objet des publications (ou pré-publications) **[B]**, **[C]**, **[D]**, **[E]**, **[I]**, **[J]** et **[K]** (voir page 6).

Dans **[A]**, nous nous sommes intéressés à une nouvelle description de l'algorithme EM très couramment utilisé en inférence selon le principe du maximum de vraisemblance pénalisée. Nous avons montré en particulier que l'algorithme EM peut se reformuler comme une méthode de type proximale avec une pénalisation de type divergence de Kullback-Leibler. Si l'algorithme EM est un peu difficile à décrire, la méthode proximale est très simple: elle consiste à générer une suite $(x^{(l)})_{l\in\mathbb{N}}$ selon la récurrence suivante:

$$x^{(l+1)} \quad \in \quad \text{argmin}_{x\in\mathcal{X}} f(x) + \lambda_l d(x; x^{(l)}), \tag{3.4}$$

où $d(\cdot;\cdot)$ est un terme de pénalisation, qui est souvent pris comme étant le carré de la norme euclidienne de $x - x^{(l)}$. L'algorithme EM est souvent utilisé dans le cas où les données sont "incomplètes". On suppose qu'on observe les données $x_1, \ldots, x_n$ alors qu'elles ne sont de partielles et que les données dites "complètes" sont du type (x_1, y_1), $\ldots$, (x_n, y_n). Un exemple simple est celui où l'on dispose de données sur les salaires de gar c cons et de filles mais que l'on ne dispose pas de la donnée des sexes dans l'échantillon. Lorsqu'on veut estimer les paramètres de la loi sous-jacente, on procède fréquemment à la maximisation de la log-vraisemblance. Dans le cas de données incomplètes, l'écriture de la log-vraisemblance est malaisée et requiert des intégrales algorithmiquement coûteuses se prêtant mal à l'incoporation dans les procédures de type Newton ou gradient conjugué pour l'optimisation. La prise en compte des données complètes même lorsqu'elles ne sont pas réellement disponibles, est une astuce technique qui permet de drastiquement simplifier l'optimisation de la vraisemblance. Soit θ^{true} le vecteur des paramètres de la loi des données, que l'on souhaiterait estimer et soit Θ un ensemble fermé auquel on sait a priori qu'il appartient. On associe, comme fréquemment, aux données numériques $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$, des variables aléatoires $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$.

Notons $l_{x_1,\ldots,x_n}(\theta)$ la log-vraisemblance des données $x_1, \ldots, x_n$ et $\mathbb{E}_\theta[\cdot]$ l'opérateur d'espérance conditionnelle selon la loi paramétrée par θ. On commence par écrire la log-vraisemblance des données complètes. Celles ci n'étant pas nécessairement observées, on s'en tire en approchant cette log-vraisemblance complète par son espérance conditionnelle sachant les données incomplètes, qui elles sont réellement observées:

$$Q\left(\theta; \theta^{(l)}\right) \quad = \quad \mathbb{E}_{\theta^{(l)}}\left[l_{(X_1,Y_1),\ldots,(X_n,Y_n)}(\theta) \mid Y_1, \ldots, Y_n\right] \tag{3.5}$$

La loi conditionnelle devrait en toute logique être spécifiée par le paramètre θ^{true} mais comme celui-ci est inconnu, on le remplace, comme dans toute méthode d'optimisation itérative, par son estimé courant, dénoté $\theta^{(l)}$ à l'itération l. On passe ensuite à l'étape de maximisation, i.e.

$$\theta^{(l+1)} \quad \in \quad \text{argmax}_{\theta\in\Theta} Q\left(\theta; \theta^{(l)}\right). \tag{3.6}$$

Supposons que la famille de densités $\{k(x|y;\theta)\}_{\theta\in\mathbb{R}^p}$ soit telle que $k(x|y;\theta)\mu(x)$ and $k(x|y;\bar{\theta})\mu(x)$ soient absolument continues pour tout θ et $\bar{\theta}$ dans Θ. Alors, la dérivée de Radon-Nikodym $\frac{k(x|y,\bar{\theta})}{k(x|y;\theta)}$ est bien définieet on peut introduire sans problème la divergence de Kullback Leibler entre les densités conditionnelles:

$$I_y(\theta,\bar{\theta}) = \mathbb{E}\big[\log\frac{k(x|y,\bar{\theta})}{k(x|y;\theta)}|y;\bar{\theta}\,\big]. \tag{3.7}$$

Définissons maintenant D_l comme le domaine de définition de l_y, $D_{I,\theta}$ comme le domaine de définition de $I_y(\cdot,\bar{\theta})$ et D_I comme celui de $I_y(\cdot,\cdot)$. On peut maintenant présenter une définition pour la famille des algorithmes de point proximal qui va servir de cadre d'étude général et flexible pour les algorithmes EM et leurs généralisations.

Definition 3.1 *Soit $(\beta_k)_{k\in\mathbb{N}}$ une suite de nombres réels positifs. Alors, l'algorithme du point proximal avec pénalisation de type Kullback-proximal algorithm est défini par*

$$\theta^{k+1} = \ argmax_{\theta\in D_l\cap D_{I,\theta^k}} l_y(\theta) - \beta_k I_y(\theta,\theta^k). \tag{3.8}$$

Le point de départ des travaux de cette partie est le résultat suivant.

Proposition 3.2 ***[B]*** *L'algorithme EM n'est rien d'autre qu'un cas particulier dans la famille des algorithmes de type point proximal avec pénalisation de Kullback-Leibler $\beta_k = 1$, pour tout $k\in\mathbb{N}$.*

La preuve de ce résultat est extrêmement naturelle. Il suffit d'écrire:

$$\theta^{k+1} = \ \text{argmax}_{\theta\in\mathbb{R}^p}\big\{\log g(y;\theta) + \mathbb{E}\big[\log\frac{f(x;\theta)}{g(y;\theta)}|y;\theta^k\big]\big\}.$$

Cette équation est équivalente à

$$\begin{aligned}\theta^{k+1} = \ \text{argmax}_{\theta\in\mathbb{R}^p}\big\{&\log g(y;\theta) + \mathbb{E}\big[\log\frac{f(x;\theta)}{g(y;\theta)}|y;\theta^k\big]\\ &- \mathbb{E}\big[\log\frac{f(x;\theta^k)}{g(y;\theta^k)}|y;\theta^k\big]\big\}\end{aligned}$$

car le terme additionnel est constant en θ. En se remémorant le fait que $k(x|y;\theta) = \frac{f(x;\theta)}{g(y;\theta)}$, on obtient

$$\begin{aligned}\theta^{k+1} = \ \text{argmax}_{\theta\in\mathbb{R}^p}\big\{&\log g(y;\theta) + \mathbb{E}\big[\log k(x|y;\theta)|y;\theta^k\big]\\ &- \mathbb{E}\big[\log k(x|y;\theta^k)|y;\theta^k\big]\big\}.\end{aligned}$$

On obtient donc finalement

$$\theta^{k+1} = \text{argmax}_{\theta\in\mathbb{R}^p}\big\{\log g(y;\theta) + \mathbb{E}\big[\log\frac{k(x|y;\theta)}{k(x|y;\theta^k)}|y;\theta^k\big]\big\}$$

ce qui conclut la preuve.

Celle formulation plus générale dans laquelle s'inscrivent les algorithmes EM va s'avérer dans la suite très intéressante pour les raisons suivantes:

- La ré-écriture sous forme Kullback-proximale permet de montrer très vite que la suite des estimées $(\theta^{(l)})_{l\in\mathbb{N}}$, augmente naturellement la log-vraisemblance à chaque itération (preuve laissée en exercice au lecteur avant de s'endormir). Ceci est un gain remarquable par rapport à la pratique courante en optimisation classique qui veut que l'on ait recourt à une procédure dite de "recherche en ligne" pour obtenir la monotonie de la suite des valeurs de la fonction objectif.
- Le cadre Kullback proximal permet d'incorporer facilement des contraintes supplémentaires sans avoir recourt à des multiplicateurs de Lagrange dont l'optimisation devrait être faite en supplément à celle des $\theta^{(l)}$. Cette potentialité sera de prime importance pour l'incorporation des contraintes de stabilité pour l'application à l'estimation de séries temporelles vectorielles de type GARCH dans la section 3.
- La suite de paramètres de relaxation $(\beta_k)_{k\in\mathbb{N}}$ permet une grande flexibilité de dans le choix des pas de la méthode et on a montré dans **[B]**, en suivant des travaux plus anciens de Rockafellar, (R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization*, vol. 14, pp. 877–898, 1976) qu'une convergence superlinéaire de la méthode pouvait être obtenue dans le cas où la limite de la suite $(\theta^{(l)})_{l\in\mathbb{N}}$ se trouvait à l'intérieur de l'ensemble Θ.

Le travail que j'ai effectué sur ces algorithmes se structure de la fa c con suivante: étude du cas concave sans contrainte, du cas de l'optimisation partielle coordonnée par coordonnée, du cas où la suite converge vers un point du bord de Θ, du cas pénalisé par une fonctionnelle non-différentiable, puis enfin l'étude de deux applications à des problèmes pratique pour l'estimation de modèles généraux additifs et le cas de l'estimation des modèles GARCH vectoriels.

Le cas concave sans contrainte

Le Théorème de convergence pour la méthode Kullback proximale de la définition 3.1 contenu dans **[B]** est le suivant.

Theorem 3.3 *Faisons les hypothèses suivantes.*

(i) $\Theta \subset \mathbb{R}^p$.

(ii) $l_y(\theta)$ *est dans* $\mathcal{C}^2$ *sur* $int\ \Theta$ *et* $I_y(\bar{\theta},\theta)$ *est dans* $\mathcal{C}^2$ *sur* $int\ \Theta \times\ int\ \Theta$.

(iii) $\lim_{\|\theta\|\to\infty} l_y(\theta) = -\infty$ *où* $\|\theta\|$ *est la norme euclidienne sur* $\mathbb{R}^p$.

(iv) $l_y(\theta) < \infty$ *et* $\Lambda_{\nabla^2 l_y(\theta)} < 0$ *sur tout borné dans* Θ.

(v) Pour tout $\bar{\theta}$ *dans* Θ, $I_y(\bar{\theta},\theta) < \infty$ *et* $0 < \lambda_{\min}\left(\nabla^2_{01} I_y(\bar{\theta},\theta)\right)$ *sur tout borné de* Θ.

(vi) L'unique maximiseur θ^* *de la vraisemblance sur* Θ *est à l'intérieur de* Θ.

Supposons la suite de paramètres de relaxation $\{\beta_k\}_{k\in\mathbb{N}}$ *positive et convergente vers* $\beta^* \in [0,\infty)$. *Alors, la suite* $\{\theta^k\}_{k\in\mathbb{N}}$ *converge vers* θ^*, *l'unique maximiseur de la log-vraisemblance* $l_{X_1,\ldots,X_n}$.

Le deuxième théorème de concerne la vitesse de convergence de la méthode pour une suite de paramètres de relaxation convergeant vers zéro. On rappelle qu'une suite $\{\theta^k\}$ est dite converger super-linéairement vers une limite θ^* si:

$$\lim_{k\to\infty} \frac{\|\theta^{k+1} - \theta^*\|}{\|\theta^k - \theta^*\|} = 0, . \tag{3.9}$$

On a alors le résultat suivant.

Theorem 3.4 *Supposons que les mêmes hypothèses que celles du Théorème 3.3 sont satisfaites et que de plus la suite* $\{\beta_k\}_{k\in\mathbb{N}}$ *converge vers zero. Alors la suite* $\{\theta^k\}_{k\in\mathbb{N}}$ *converge superlinéairement vers* θ^**, l'unique maximiseur de la log-vraisemblance* $l_{X_1,\ldots,X_n}$.

Une remarque amusante pour conclure: le cas de l'estimation pour les mélanges de gaussiennes, qui est l'instance où l'algorithme EM est le plus utilisé, ne vérifie pas les conditions de ces théorèmes. Les problèmes suivants sont extrêmement délicats à prendre en compte.

- La log-vraisemblance n'est pas bornée et il est bien connu que des solutions dégénérées existent, c'est à dire des solutions correspondant à une composante du mélange caractérisée par une matrice de covariance de rang défectueux. Ce problème est très important, dans philosophiquement qu'en pratique. Cela signifie en particulier que la maximisation de la vraisemblance n'est pas pertinente pour ce type de problème, même si en pratique, des filouteries ad hoc permettent de s'en sortir. Nous discuterons plus précisément de ces aspects dans la Section 3;

- La log-vraisemblance n'est pas concave. Une des manières de le voir est la considération très naturelle du problème de "label switching" qui dit que si θ^* est un maximizeur, la permutation des indices spécifiant le numéro de chaque composante donne un nouveau maximiseur. De plus, comme le problème de dégénérescence l'indique, la suite doit pouvoir converger vers le bord de Θ, un cas qui n'est pas considéré dans ce premier travail; ces problèmes seront adressés dans la Section 3.

- La distance de Kullback-Leibler $I_y(\theta;\bar{\theta})$ entre deux densités $k(\cdot;\theta)$ et $k(\cdot;\bar{\theta})$ n'est pas satisfaisante dans le sens où $I_y(\theta;\bar{\theta}) = 0$ n'implique pas $\theta = \bar{\theta}$. Ce dernier problème est examiné plus dans le travail que je présente dans la section suivante.

Le cas des mélanges gaussiens (1): utilisation de la maximisation coordonnée par coordonnée

Le problème suppose que l'on observe des données mutlidimensionnelles $x_1,\ldots,x_n \in \mathbb{R}^d$ provenant d'une densité de mélange de gaussiennes donnée par

$$f_X(x) \quad = \quad \sum_{j=1}^{J} \pi_j \phi_j(x), \tag{3.10}$$

où ϕ_j est la densité gaussienne donnée par

$$\phi_j(x) \quad = \quad \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_j)}} \exp\left(-\frac{1}{2}(x-\mu_j)^t \Sigma_j^{-1}(x-\mu_j)\right). \tag{3.11}$$

L'algorithme CEMM considère une décomposition du vecteur de paramètres $\theta = (\theta_j, j = 1,...,J)$ avec $\theta_j = (\pi_j, \mu_j, \Sigma_j)$. et choisit de mettre à jour une coordonnée arbitraire à chaque itération plutôt que toutes les coordonnées d'un coup. Par simplicité, nous nous restreindrons dans la suite à une présentation dans le cas où les indices des coordonnées mises a jour sont choisis cycliquement, en commen c cant par $j = 1,\ldots,J$ et en répétant cette séquence toutes les J itérations. Ainsi, la composante mise à jour à l'itération k est $j = k - \frac{k}{J}\rfloor J + 1$, où $.\rfloor$ dénote la partie entière.

La maximisation coordonnée par coordonnée peut être préférée pour plusieurs raisons. Dans le cas des mélanges gaussiens, nous avons constaté que la convergence était vraiment

améliorée en pratique. Ceci en fait une approche très appréciée et l'article **[C]** est le plus cité parmi les travaux que je présente dans ce document. L'explication du phénomène de l'accélération de la convergence est certainement à chercher dans l'article de Hero et Fessler (Fessler, J. A., and Hero, A. O. (1994), "Space-Alternating generalized expectation-maximisation algorithm", *IEEE Trans. Signal Processing*, 42, 2664-2677) où une analyse géométrique de la vitesse en fonction de la courbure de la log-vraisemblance selon la coordonnée (ou plus généralement, le sous-espace) sélectionné(e) est fournie. Spécifier cette théorie dans le cas des mélanges semble encore un peu trop difficile pour pouvoir en exploiter toutes les finesses et c'est pour cette raison que nous nous sommes intéressés à d'autres questions. La seconde raison pour laquelle on peut préférer l'approche par coordonnée est que dans ce cas, $I_y(\theta;\bar{\theta}) = 0$ pour deux vecteurs θ et $\bar{\theta}$ ne différant que sur une composante implique que $\theta = \bar{\theta}$, ce qui, comme nous l'avons déjà souligné dans la section précédente, n'est pas le cas pour θ et $\bar{\theta}$ généraux, comme cela semble être ignoré couramment. Même si les parties de l'article pionnier de Dempster Laird et Rubin concernant les propriétés théoriques de convergence semblent montrer quelques faiblesses (corrigées successivement pendant les années qui suivirent par plusieurs auteurs dont Wu (Wu, C. F. (1983), On the convergence of the EM algorithm, Annals of Statistics, 11, 95-103)), il y est tout de même préconisé de verifier la propriété que deux itérés successifs se rapprochent asymptotiquement. Or sans la condition "$I_y(\theta;\bar{\theta}) = 0$ implique $\theta = \bar{\theta}$", comme dans le cas des mélanges, il parait difficile de le démontrer. L'approche composante par composante permet de prouver rigoureusement ce type de résultat naturel. On obtient même le résultat suivant, qui est le théorème principal concernant la convergence de l'algorithme CEMM.

Theorem 3.5 *Chaque point d'accumulation de* $\left\{\theta^k\right\}_{k\in\mathbb{N}}$*, pour lequel toutes les matrices de covariances sont de rang plein, est un point stationnaire de la log-vraisemblance sous la contrainte* $\sum_{\ell=1}^{J} \pi_\ell = 1$.

Le cas des mélanges gaussiens (2): dégénérescence de l'algorithme EM

Comme déjà mentionné dans la présentation de l'algorithme EM, la méthode peut générer une suite d'itérés convergeant vers un point du bord de Θ. Dans le cas des mélanges gaussiens, une telle situation arrive lorsque l'une des composantes d'un point d'accumulation est dégénérée, c'est à dire que la matrice de covariance associée est singulière. Nous avons étudié le cas particulier de la dimension un avec Christophe Biernacki dans **[D]** et démontré le résultat suivant.

Theorem 3.6 *Soit* $f_{i,k} = \pi_k \phi(x_i;\mu_k,\sigma_k^2)$ *et soit* u_0 *le vecteur dont les composantes sont* $1/f_{i_0,k_0}$ *et* f_{i,k_0}*,* $i \neq i_0$*. Il existe* $\varepsilon > 0$*,* $\alpha > 0$ *et* $\beta > 0$ *tels que si* $\|u_0\|_2 \leq \varepsilon$*, alors*

$$\sigma_{k_0}^{2+} \quad \leq \quad \alpha \frac{\exp\left(-\beta/\sigma_{k_0}^2\right)}{\sigma_{k_0}^2}. \tag{3.12}$$

Ainsi, si par malheur on est proche d'une situation critique, la variance tend extrêmement rapidement vers 0. L'idée est qu'il vaut mieux laisser l'algorithme dériver vers une telle dégénérescence car elle est finalement très rapidement détectable. Une approche bayésienne consistant à pénaliser la distance à la singularité des matrices de covariance semble un peu effrayante a posteriori: pénaliser pour éviter la dégénérescence peut tout simplement créer un maximiseur artificiel où l'algorithme risque de s'enliser sans qu'on sache qu'il s'y trouve. S'écraser rapidement sur un point dégénéré permet de se rendre compte qu'on se dirigeait

vers un point absurde. Il suffit alors de relancer l'algorithme EM à partir d'un nouveau point initial jusqu'à ce que l'on constate la convergence vers un point plus raisonnable. Ceci ne résoud pas la question de savoir si un tel point "plus raisonnable" est effectivement pertinent, étant donné l'existance de maximiseur non-globaux à la vraisemblance, mais nous permet déjà de mieux comprendre le comportement de l'algorithme EM pour un problème très important et de ne pas introduire de nouvelle bêtise dans son utilisation.

Le cas d'une log-vraisemblance non-convexe et de points d'accumulation sur le bord

Dans l'article [**E**] écrit avec Alfred Hero, je me suis intéressé au cas plus général de log-vraisemblances non-convexes et de point d'accumulation sur le bord. La non-convexité est importante à prendre en compte car elle se rencontre très souvent dans ce domaine, avec comme exemple important celui des mélanges de gaussiennes. Par contre, l'adaptation théorique est très facile et on ne peut pratiquement rien prouver d'intéressant: les points d'accumulation à l'intérieur de Θ ne sont rien d'autre a priori que des points stationnaire de la log-vraisemblance, c'est à dire des points qui annulent le gradient. Ces points peuvent donc être des maximiseurs locaux, des minimiseurs locaux ou encore des points selles. On est bien avancés ! Il se trouve malgré tout que de pouvoir démontrer ce type de propriété est déjà considéré comme un réjouissement étant donné l'écart qui reste devant l'accès à des propriétés plus satisfaisante comme celle d'être un maximiseur global par exemple. Le cas où les points stationnaires sont sur le bord de Θ est quand même un peu plus délicat à étudier que le cas où ils sont à l'intérieur. Cela vient du fait que la pénalisation $I_y(\theta;\bar{\theta})$ "explose" sur le bord de Θ. Exploser signifie ici que $\lim_{\theta\to\theta^*} I_y(\theta,\bar{\theta}) = +\infty$ quel que soit $\bar{\theta} \in \text{int}\Theta$. Il n'y a que lorsque l'on a conjointement, et de manière équilibrée, rapprochement asymptotique de $\theta^{(l)}$ vers $\theta^{(l+1)}$ et rapprochement de $\theta^{(l)}$ vers le bord de Θ, que l'on peut espérer eviter l'explosion. Pour un problème donné, on peut souvent montrer simplement que cette explosion n'arrive pas. Cela doit donc vouloir dire que deux itérés successifs se rapprochent tout en se rapprochant du bord, mais peut-on facilement garantir qu'un point d'accumulation de cette suite satisfait des conditions nécéssaires rudimentaires d'optimalité ? Dans [**D**], je démontre que c'est le cas, en ce sens que les points d'accumulation sur le bord de Θ satisfont bien les conditions de Karush-Kuhn-Tucker.

Le cas d'une log-vraisemblance non-convexe avec une pénalisation non-différentiable

Dans [**I**], nous continuons l'étude présentée dans [**E**] dans le cas où on ajoute une pénalisation non-différentiable et on maximise à chaque itération sur la restriction à un sous-espace choisi arbitrairement. L'étude se passe presque mot pour mot comme pour [**E**] sauf que la non-différentiabilité de la pénalisation crée des interférences avec le fait qu'on maximize relativement à des sous-espaces et l'on n'est pas sur d'obtenir à la fin un point stationnaire verifiant les conditions de Karush-Kuhn-Tucker (généralisée au sens des sous-gradients de Clarke). L'intérêt de la pénalisation non-différentiable est de pouvoir dans certaines conditions promouvoir la parcimonie de la solution. L'intérêt de pouvoir optimiser le long de sous-espaces est de rendre facile à implanter certaines versions de l'algorithme EM un peu lourdes.

Une application est développée dans [**J**] où je propose de résoudre le problème de l'estimation d'un mélange de gaussiennes dans le cas où peu d'observations sont disponibles et la dimension du problème peut être relativement élevée. L'idée est de postuler que les gaussiennes ont pour espérance une combinaison sparse des données. On écrit alors la vraisemblance en postulant que les espérances sont une régression des données elles-mêmes.

La parcimonie est obtenue par pénalisation de la norme ℓ_1 du vecteur de régression. La méthode est décrite dans le cas de variances toutes multiples de l'identité, ce qui n'est pas choquant quand peu de données sont disponibles, une situation où il est nécessaire de diminuer au maximum le nombre de paramètres à estimer. Il est cependant très facile de mettre en oeuvre la méthode dans le cas de matrices de covariances quelconques ou structurées. Les expériences de simulations montre que la méthode est assez performante en comparaison de EM classique pour les mélanges de gaussiennes et CEM (Classification EM).

Application des méthodes Bregman Proximales locales à l'estimation de modèles GARCH vectoriels

Dans [**M**] nous avons, avec Juan Pablo Ortega, appliqué les techniques proximales étudiées précédemment dans le cas de l'estimation par maximum de vraisemblance pour les modèles de séries temporelles vectorielles à volatilité gouvernée par un modèle GARCH. Ce problème était reconnu comme numériquement très difficile. Un des points très délicats était notamment la prise en compte des conditions de stationarité du modèle. Il fut d'ailleurs très cocasse de constaté lors de l'exploration méticuleuse de la littérature sur le sujet, que les conditions n'étaient jamais très facilement exploitable d'un point de vue numérique (voir par exemple l'ouvrage connu de Gourieroux sur le sujet). Nous avons alors introduit des opérateurs permettant de décrire proprement des conditions suffisantes assurant que le processus de covariance conditionnel matriciel soit bien positif semi défini et assurant aussi par ailleur l'existance d'une solution stationnaire. Nous avons ensuite appliqué une approche de type Bregman proximale pour optimiser la vraisemblance sous les contraintes spectrales que nous avons introduites. Cet algorithme est robuste et rapide en comparaison avec les autres approches que nous avons tentées pour ce problème délicat. Il est aussi le seul à maximiser la vraisemblance pour des modèles GARCH aussi généraux sous les contraintes nécéssaires à la bonne définition de la série temporelle. Entrons maintenant dans les détails.

Un processus $\{\mathbf{z}_t\}$ à temps discret n-dimensionnel est dit conditionnellement heteroscedastique s'il est déterminé par les relations

$$\mathbf{z}_t = H_t^{1/2}\boldsymbol{\varepsilon_t} \qquad \text{with} \qquad \{\boldsymbol{\varepsilon_t}\} \sim \text{ IIDN}(\mathbf{0}, \boldsymbol{I}_n).$$

Dans cette expression, $\{H_t\}$ denote un processus matriciel prévisible, c'est à dire que pour tout $t \in \mathbb{N}$, la variable aléatoire matricielle H_t est $\mathcal{F}_{t-1}$-mesurable, et $H_t^{1/2}$ est la racine carrée de H_t, c'est à dire $H_t^{1/2}(H_t^{1/2})^T = H_t$. Il est facile de voir que l'espérance conditionnelle satisfait $E_t[z_t] = \mathbf{0}$ et que le processus de matrices de covariance conditionnelle $\{z_t\}$ est donné par $\{H_t\}$.

Le modèle VEC-GARCH (ou simplement VEC par abus de langage) a été introduit par Bollerslev comme une généralisation directe du modèle GARCH unidimensionnel (T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, Journal of Econometrics, 31 (1986), no. 3, 307–327) dans le sens où toutes les covariances conditionnelles sont des fonctions des covariances conditionnelles aux instants passés ainsi que des produits croisés des observations. Plus précisément, le modèle VEC(q,p) est donné par

$$\boldsymbol{h}_t = c + \sum_{i=1}^{q} A_i \boldsymbol{\eta}_{t-i} + \sum_{i=1}^{p} B_i \boldsymbol{h}_{t-i},$$

où $\boldsymbol{h}_t :=$ vech(H_t), $\boldsymbol{\eta}_t := z_t z^T$, c est un vecteur N-dimensionnel, avec $N := n(n+1)/2$ et $A_i, B_i \in \mathbb{M}_N$.

Nous nous sommes concentrés sur le cas $p = q = 1$, car le modèle général est facilement obtenu à partir de ce cas simple:

$$\begin{cases} \mathbf{z}_t &= H_t^{1/2}\boldsymbol{\varepsilon_t} \quad \text{with} \quad \{\boldsymbol{\varepsilon_t}\} \sim \text{IIDN}(\mathbf{0}, \boldsymbol{I}_n), \\ \boldsymbol{h}_t &= c + A\boldsymbol{\eta}_{t-1} + B\boldsymbol{h}_{t-1}. \end{cases} \tag{3.13}$$

Pour ce modèle simple, on a déjà besoin de $N(2N+1) = \frac{1}{2}(n^2+n)(n^2+n+1)$ paramètres pour spécifier complètement les objets.

La description générale que nous venons de présenter du modèle VEC (3.15) ne nous garantit pas qu'il existe une solution stationnaire. De plus des contraintes additionnelles sur c, A, et B afin de s'assurer que le processus $\{H_t\}_{t\in\mathbb{N}}$ soit positif semi-défini. Malheureusement, il semble difficile de trouver de telles contraintes sous forme aisément implémentable dans les ouvrages classique (comme celui de Gourieroux), et nous avons préférer introduire nos propres conditions suffisantes qui se sont avéré très bien adaptées à l'optimisation numérique.

Constraintes de positivité: Nous avons établi la propriété très pratique suivante.

Proposition 3.7 *Si les paramètres c, A, et B dans (3.15) sont tels que $math(c), \Sigma(A)$, et $\Sigma(B)$ sont positives semi-définie, alors les matrices de covariance conditionnelles $\{H_t\}_{t\in\mathbb{N}}$ le sont aussi si H_0 l'est déjà elle même.*

Constrainte de stationnarité du second ordre: Gourieroux a donné des conditions en termes de rayon spectral de $A + B$. Ces conditions étant difficile à incorporer à un scheme d'optimisation, nous avons choisi de les rendre un peu plus contraignantes sur la base de la proposition suivante.

Proposition 3.8 *Un modèle VEC spécifié comme dans (3.15) admet une unique solution stationnaire au second ordre si toutes les valeurs singulières de $A + B$ sont dans le cercle unité ouvert. C'est en particulier toujours le cas dès que $\sigma_{max}(A+B)$ est plus petite que un, où encore, dès que $\mathbb{I}_N - (A+B)(A+B)^T$ est positive semi-définie. Si une de ces conditions est satisfaite, alors*

$$\Gamma(0) = math(E[h_t]) = math((\mathbb{I}_N - A - B)^{-1}c). \tag{3.14}$$

Abordons maintenant l'optimisation proprement dite en intruisant la divergence de Bregman et l'algorithme proximal associé. La divergence de Bregman matricielle est définie comme suit.

Definition 3.9 *Soient X, Y des matrices symétriques réelles d'ordre n et ϕ une fonction strictement convexe différentiable. La Divergence matricielle de Bregman associée à ϕ est définie par*

$$D_\phi(X, Y) := \phi(X) - \phi(Y) - \text{trace}\left(\nabla\phi(Y)^T(X - Y)\right).$$

Les divergences de Bregman sont utilisées pour mesurer la proximité entre deux matrices. En particulier, si $\phi(X) := \|X\|^2$, alors, $D_\phi(X, Y) := \|X - Y\|^2$. Un autre exemple est la divergence de von Neumann qui n'est rien d'autre que la divergence de Bregman associée à l'entropie; Plus spécifiquement, si X est une matrice symétrique réelle positive semi-définie de valeurs propres $\{\lambda_1, \ldots, \lambda_n\}$, alors $\phi(X) := \sum_{i=1}^n(\lambda_i \log \lambda_i - \lambda_i)$. Dans notre implémentation, nous avons utilisé la divergence de Brug (encore appelée divergence LogDet où fonction de perte de Stein dans la litérature statistique) qui est la divergence de Bregman obtenue avec le choix $\phi(X) := -\sum_{i=1}^n \log \lambda_i$, ou de manière équivalente

$\phi(X) := -\log\det(X)$. La divergence de Bregman qui est résulte sur les matrices positives définies est donnée par

$$D_B(X,Y) := \mathrm{trace}(XY^{-1}) - \log\det(XY^{-1}) - n. \tag{3.15}$$

Supposons maintenant qu'on veuille résoudre le problème d'optimisation suivant

$$\arg\min_{A\succeq 0} f(A),$$

sous la contrainte $A \succeq 0$. Une fa c con de procéder consiste à optimiser un modèle local, pénalisé par une divergence de Bregman:

$$f_{A^{(n)}}(A) := f\left(A^{(n)}\right) + \left\langle \nabla f\left(A^{(n)}\right), A - A^{(n)} \right\rangle + \frac{L}{2} D_\phi(A, A^{(n)}). \tag{3.16}$$

L'intérêt d'introduire la divergence de Brug, est que la solution à chacun de ces sous problèmes locaux est elle même positive définie et que les points d'accumulation de la suite engendrée par l'approche ne peut être, s'ils existent, que positif semi-défini. Nous renvoyons le lecteur à l'article **[I]** pour les détails de l'implémentation permettant de prendre en compte la vraisemblances et les contraintes de positivité ainsi que de stationnarité.

Application d'EM à un mélange de modèles GAM et application à la construction d'une carte de prévalence

Avec David Pleydell, du Laboratoire de Biologie Environnementale, maintenant incorporé à la structure Chrono-eco-environnement, nous nous sommes penché sur l'estimation d'un certain type de modèle GAM pour la prévision de la prévalence sur une zone géographique de la Chine proche du Tibet. Cette étude faisait partie d'un grand projet financé par la NIH et dont le but était de comprendres les facteurs d'infection par l'echinococose alvéolaire. Nous avons tout simplement implanté un algorithme EM pour un modèle de mélange de modèles additifs et l'avons mis en pratique sur les données récoltée par les équipes du laboratoire de chrono-environnement.

Les modèles de type Generalised additive models (GAMs) sont devenus récemment très populaires en écologie de part leur capacités à prendre en compte les possibles non-linéarités. L'approche usuelle consiste à ajouter des fonctions lisses des covariables dans le prédicteur du modèle linéaire généralisé. On a choisi en particulier

$$g(\mu_i) = \beta_0 + \beta_1 \mathcal{H}(x_i)$$

avec $\mu \equiv E[Y]$, où Y suit une distribution dans la famille exponentielle, β_0 est l'ordonnée à l'origine et $\mathcal{H}$ est une fonction lisse de la covariable x. Le choix le plus courant pour la fonction $\mathcal{H}$ est de prendre une fonction spline car ce type de fonction est extrêmement flexible. Dans ce travail, nous avons fait un choix beaucoup plus simple et mieux adapté à l'application biologique: dans notre modèle GAM $\mathcal{H}$ est définie comme l'application unimodale

$$\mathcal{H}_{\alpha_1,\alpha_2}(x) = \frac{\left(\frac{x-l}{u-l}\right)^{\alpha_1}\left(\frac{u-x}{u-l}\right)^{\alpha_2}}{\left(\frac{m-l}{u-l}\right)^{\alpha_1}\left(\frac{u-m}{u-l}\right)^{\alpha_2}}.$$

Cette transformation va permettre la détection des valeurs de x correspondant aux zones où l'espèce biologique étudiée sera plus encline à former une niche écologique. Les paramètres α_1 et α_2 peuvent prendre leurs valeurs dans $(0,\infty)$.

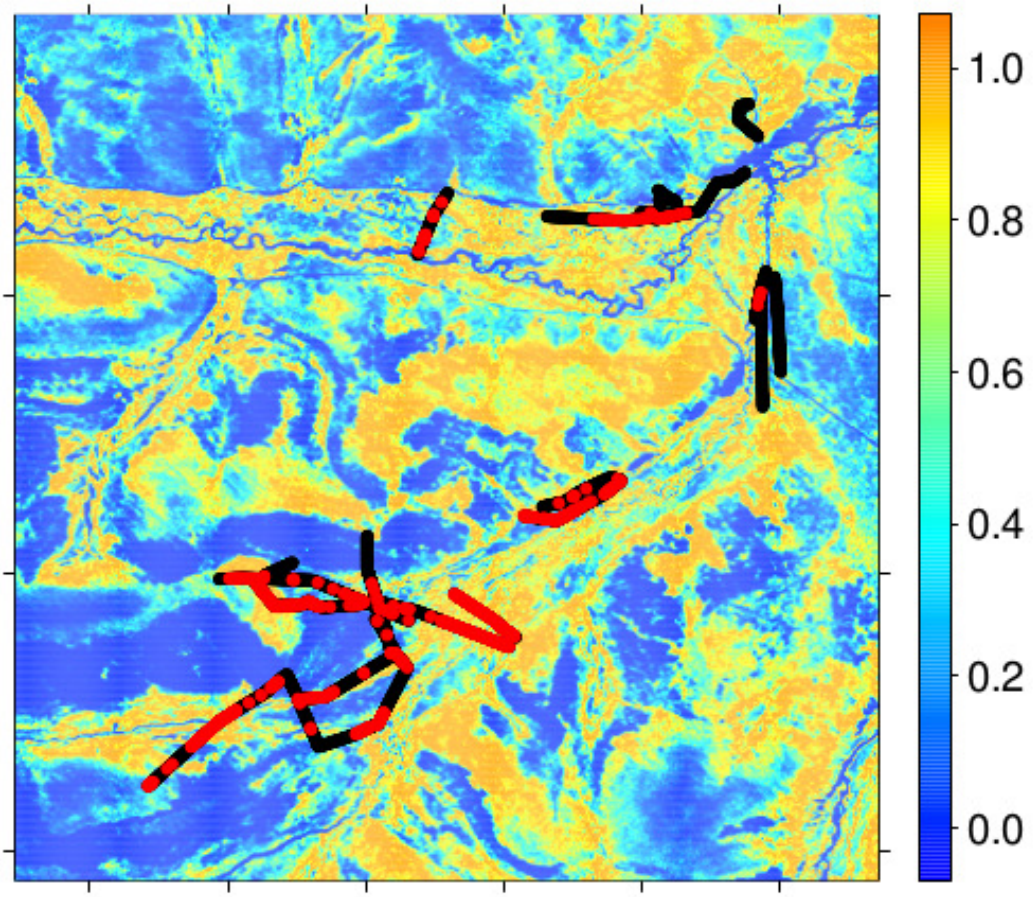

Figure .1: Habitat Suitability Index derived from the NDVI using ML estimates of $\hat{\alpha}$ and $\hat{m}_1$ from $\mathcal{M}_3$ overlayed with transect data on small mammal indices. Red and black points repressent pressence and absence of observable small mammal indeced respectively.

Avec ce type d'approche, on a pu donner des cartes de l'indice de pertinence pour l'habitat dont la Figure F.3 est un exemple, ainsi que d'autres éléments biologiques pour lesquels on renvoit le lecteur à l'article.

Le travail n'a pas généré de difficulté mathématique particulière, mis à part toutes les questions classiques que l'on se pose sur les modèles de mélange et la convergence de l'algorithme dans les situations où la log-vraisemblance n'est pas convexe. La mise en oeuvre fut un peu laborieuse mais s'est mise à fonctionner après quelques mois de méticuleuses corrections dans les moments de motivation exarcerbée. Le fait que l'implantation de tels modèles demande beaucoup de soin et de patience lui a valu de pouvoir être publié dans un journal internationnal à comité de lecture [**G**]. Je renvoie le lecteur à l'article pour plus de détails.

4 Le Compressed Sensing, le LASSO et les matrices aléatoires

La deuxième partie du travail post-thèse est la plus récente et marque une réorientation vers des problèmes plus à la mode concernant les matrices aléatoires et le Compressed Sensing. Le Compressed Sensing est un domaine qui a récemment émergé suite à une longue exploration que l'on peut identifier comme issue de l'analyse par ondelettes créée par Morlet, Grossmann, Meyer, Daubechies, Mallat, Donoho, Johnstone et leur collaborateurs, dans les années 80 et l'approximation finie de cette analyse. Nous renvoyons le lecteur à l'ouvrage A Mathematical Introduction to Wavelets, par P. Wojtaszczyk pour une introduction très pédagogique aux ondelettes et aux propriétés des décompositions selon les espaces fonctionnels auxquelles la fonction décomposée appartient. Une méthode possible pour approcher une fonction de $L^2(\mathbb{R}^d)$ est de la décomposer en ondelettes, i.e.

$$f = \sum_{k\in\mathbb{Z}} c_k\phi_k + \sum_{j=1}^{\infty}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k} \tag{4.17}$$

où $\phi_k(\cdot) = \phi(\cdot - k)$ et $\psi_{j,k}(\cdot) = 2^{j/2}\psi(2^j \cdot -k)$, ϕ est la fonction d'échelle et ψ est une ondelette associée, puis de conserver les termes correspondant à toutes les échelles jusqu'à une échelle maximale $J_{\max}$. On peut aussi se contenter des termes $d_{j,k}$ correspondant à des translations petites (k dans un ensemble $\mathcal{K}_j$ par exemple), ce qui est naturel lorsqu'on observe une fonction sur un intervalle compact. On obtient ainsi l'approximation

$$A_s^L(f) = \sum_{k\in\mathcal{K}_0} c_k\phi_k + \sum_{j=1}^{J_{\max}}\sum_{k\in\mathcal{K}_j} d_{j,k}\psi_{j,k} \tag{4.18}$$

où s est le nombre total de termes conservés. Ce type d'approche est linéaire (d'où le superscript "L") car l'operateur A_s^L est linéaire et les indices des coefficients conservés ne dépendent pas de la fonction f.

Une stratégie plus adéquate pour approximer une fonction est la méthode non-linéaire consistant à ne conserver que les coefficients c_k et $d_{j,k}$ correspondant aux s plus grandes valeurs de la norme de $c_k\phi_k$ et $d_{j,k}\psi_{j,k}$ dans L^p respectivement. On obtient alors une approximation

$$A_s^{NL}(f) = \sum_{k\in\mathcal{K}_0(f)} c_k\phi_k + \sum_{j=1}^{J_{\max}(f)}\sum_{k\in\mathcal{K}_j(f)} d_{j,k}\psi_{j,k} \tag{4.19}$$

de f à l'aide de s termes qui semblent particlièrement bien choisis. Cette approche est non-linéaire car le choix des coefficients à conserver dépend de la fonction f. Les performances de ce type d'approximation non-linéaires ont été étudiées dans l'article R. DeVore, B. Jawerth, et V. Popov, (1992). Compression of wavelet decompositions. American Journal of Math, 114, 737–285. Un des résultats principaux de cet article est le fait que

$$\|f - A_s^{NL}(f)\|_{L_p} \sim s^{-r/d} \tag{4.20}$$

pour toute fonction f dans l'espace de Besov $B^r_{q,q}$ avec $1/q = 1/p + r/d$. D'autre part, on peut aussi démontrer qu'une telle vitesse d'approximation n'est pas possible pour toutes les fonctions de $B^r_{q,q}$ avec une méthode linéaire.

De ces études sont probablement nées les idées que certaines fonctions (signaux, images, etc ...) sont facilement approximables par s termes de leurs décompositions en ondelettes bien choisis. L'idée que beaucoup de signaux sont compressibles à alors commencé à circuler

couramment dans la communauté du traitement du signal et des images. On en est venue alors à considérer que ces signaux étaient grosso modo s-sparses dans la base d'ondelettes considérée. Dans le cas de bases orthogonales et en considérant la norme de L_2 il est facile de voir qu'il suffit pour approcher non-linéairement la fonction, de conserver les s coefficients les plus grands. Cette méthode s'appelle le seuillage fort. Une autre méthode consiste à appliquer un seuillage dit seuillage doux et qui consiste à enlever une quantité τ fixée à chaque coefficient positif et de le mettre à zéro si le résultat de cette opération est négatif et symétriquement pour les coefficients négatifs, à ajouter τ et à le mettre à zéro si le résultat est positif. Il est alors apparu que ce seuillage doux correspondait à la solution d'un problème de minimization:

$$\min_{(c_k,d_{j,k})_{j=1,\dots,+\infty,k\in\mathbb{Z}}} \left\| f - \sum_{k\in\mathbb{Z}} c_k\phi_k + \sum_{j=1}^{+\infty}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k} \right\|_{L_2}^2 + \lambda \left\|(c_k,d_{j,k})\right\|_{\ell_1} \tag{4.21}$$

pour λ bien choisi.

Puis les bases sont devenues obsolètes et on a préféré utiliser des frames ou des concaténations de bases. Le système dans lequel on a commencé à représenter les images et les signaux étant redondants, on a appelé ces systèmes des "dictionnaires". On écrit tout simplement la représentation de f dans un dictionnaire Φ indicé par $\mathcal{K}$, comme

$$f \quad = \quad \sum_{k\in\mathcal{K}} c_k\phi_k. \tag{4.22}$$

D'un autre coté, on a conservé l'approche issue du problème d'optimisation (4.23) pour obtenir une approximation non-linéaire de f et l'étude de l'efficacité d'une telle approche à engendré une multitude de travaux.

L'idée du Compressed Sensing est apparue en 2004 suite à des travaux de Candès et Romberg sur le problème inverse en imagerie médicale, et plus précisément pour l'IRM. Ils ont en effet constaté numériquement que lorsqu'on avait une fonction f exactement s sparse dans un dictionnaire ayant de bonnes propriétés, alors la solution du problème

$$\min_{(c_k)_{k\in\mathcal{K}}} \left\|(c_k)\right\|_{\ell_1} \textbf{ sous la contrainte } \quad f = \sum_{k\in\mathcal{K}} c_k\phi_k \tag{4.23}$$

était un vecteur s-sparse qui retrouvait exactement la décomposition de f dans le dictionnaire ϕ. Bien sur le cas de leur étude numérique était une situation discrétisée où f est en fait un vecteur de $\mathbb{R}^n$, le dictionnaire Φ est représenté par une matrice dans $\mathbb{R}^{n\times p}$ et la décomposition de f dans le dictionnaire Φ s'écrit

$$f \quad = \quad \Phi c, \tag{4.24}$$

où c est le vecteur de coordonnées les c_k. Ce résultat à été considéré comme très impressionnant pour une raison très simple, fondée sur des considérations de complexité algorithmiques. Supposons p très grand devant n, c'est à dire que le dictionnaire est vraiment très redondant. Supposons aussi que s est plus petit que $n/2$ et représente le cardinal du plus petit ensemble d'indices S tel qu'il existe un vecteur c^* de support S satisfaisant (4.24). Alors un lemme très facile (voir par exemple Cohen, Dahmen et DeVore, Compressed Sensing and best k-term approximation, Journal of the American Mathematical Society, vol. 22 (2009), no. 1, 211–231) dit que c^* est l'unique solution du problème d'optimisation

$$\min_c \left\|c\right\|_{\ell_0} \quad \text{sous la contrainte } f = \Phi c. \tag{4.25}$$

où $\|\cdot\|_{\ell_0}$ denote la taille du support du vecteur $\cdot$, une fonction très fortement non-convexe. Les résultats de Candès et Romberg ne disent rien de moins que ce problème peut être parfois résolu en rempla c cant $\|\cdot\|_{\ell_0}$ par $\|\cdot\|_{\ell_1}$, c'est à dire en résolvant

$$\min_c \|c\|_{\ell_1} \quad \text{sous la contrainte } f = \Phi c. \tag{4.26}$$

D'un autre coté, on sait que le problème (4.25) est NP-difficile, alors que le problème (4.26) est un problème de complexité polynomiale, ce qui peut se voir très simplement en le réduisant à un problème de programmation linéaire.

Quelles sont les conditions sur Φ permettant ce miracle ? Depuis les premiers travaux de Candès Romberg et Tao, Donoho et bien d'autres, les choses se sont un petit peut éclaircies. Il a "suffit" d'introduire la condition de quasi-Isométrie Restreinte (Restricted (quasi-)Isometry Property=RIP). Plus précisément, on dit que la matrice Φ satisfait la condition $RIP(\delta, k)$ si

$$(1-\delta)\|x\|_2 \quad \leq \|\Phi x\|_2 \leq \quad (1+\delta)\|x\|_2 \tag{4.27}$$

pour tout x dans $\mathbb{R}^p$ qui soit également k-sparse. Si la matrice Φ satisfait une telle condition avec $k = 2s$ et δ suffisamment petit, alors, on peut démontrer qu'elle satisfait la propriété suivante, dite "Null Space Property", $(NSP(s, C))$

$$\|h_T\|_{\ell_1} \quad \leq \quad C\|h_{T^c}\|_1 \tag{4.28}$$

pour tout h dans le noyau de Φ, pour tout $T \subset \{1, \ldots, p\}$ de cardinal s, et pour C une constante plus petite que 1. (La notation h_T denote ici la restriction de h à ses composantes indicées par T.) A partir de la propriété $NSP(s, C)$, on prouve alors facilement que la solution de (4.26) est le vecteur c^*. On procède de la fa c con suivante. Notons c° une solution de (4.26). On a alors

$$\|c^\circ\|_{\ell_1} \quad \leq \quad \|c^*\|_{\ell_1}. \tag{4.29}$$

Notons $h = c^\circ - c^*$. Alors, h est dans le noyau de Φ par définition de c° et c^*. On décompose alors $\|c^\circ\|_{\ell_1}$ comme suit:

$$\|c^\circ\|_{\ell_1} \quad = \quad \|c^*_S + h_S\|_{\ell_1} + \|c^*_{S^c} + h_{S^c}\|_{\ell_1} \tag{4.30}$$

où je rappelle que S est le support de c^*. Ainsi,

$$\|c^\circ\|_{\ell_1} \quad = \quad \|c^*_S + h_S\|_{\ell_1} + \|h_{S^c}\|_{\ell_1} \tag{4.31}$$

et par application de l'inégalité triangulaire, on obtient que

$$\|c^\circ\|_{\ell_1} \quad \geq \quad \|c^*_S\|_{\ell_1} - \|h_S\|_{\ell_1} + \|h_{S^c}\|_{\ell_1}. \tag{4.32}$$

En combinant cette dernière inégalité avec (4.29), on a

$$\|c^*_S\|_{\ell_1} - \|h_S\|_{\ell_1} + \|h_{S^c}\|_{\ell_1} \quad \leq \quad \|c^*_S\|_{\ell_1} \tag{4.33}$$

ce qui donne

$$\|h_{S^c}\|_{\ell_1} \quad \leq \quad \|h_S\|_{\ell_1}. \tag{4.34}$$

La propriété $NSP(s, C)$ implique alors

$$\|h_{S^c}\|_{\ell_1} \quad \leq \quad C\|h_{S^c}\|_{\ell_1} \tag{4.35}$$

avec $C \in (0,1)$ et donc $\|h_{S^c}\|_{\ell_1} = 0$ et $\|h_S\|_{\ell_1} = 0$ ce qui donne finalement

$$c^\circ \quad = \quad c^*. \tag{4.36}$$

Une fois accordés sur le fait que la condition RIP etait la clé de la compréhension du problème de retrouver un vecteur sparse à partir d'un faible nombre d'observations s'est posée la question de savoir si on pouvait facilement construire des dictionnaires (i.e. dans le cas discrétisé, des matrices) Φ vérifiant cette propriété. Il est en fait facile de voir d'après les récents travaux sur les matrices aléatoires que si on prend Φ comme étant une matrice dont les composantes sont indépendantes, identiquement distribuées et de loi gaussienne $\mathcal{N}(0, \frac{1}{\sqrt{n}})$, alors la conditon RIP est satisfaite avec forte probabilité pour s inférieur à une quantité de l'ordre de $n/\log(p/n)$. Ce résultat démontre donc que si on choisit bien le dictionnaire Φ, on peut retrouver un vecteur sparse avec un nombre d'observation de l'ordre de grandeur égal à la sparsité, à un facteur correctif de l'ordre du log de p, ce qui semble vraiment très peu cher à payer pour résoudre un problème sensé être NP-dur dans toute sa généralité. Ce résultat est en fait assez universel car il tient aux constantes près pour toutes les lois sous-gaussiennes, la loi de Bernoulli ± 1 incluse évidemment. On pourra lire avec grand intérêt les notes de cours de Vershynin sur les matrices aléatoires de taille finie: R. Vershynin, Introduction to the non-asymptotic analysis of random matrices. Chapter 5 of the book Compressed Sensing, Theory and Applications, ed. Y. Eldar and G. Kutyniok. Cambridge University Press, 2012. pp. 210–268.

Cette théorie a ensuite été étendue au cas où f n'est pas exactement observée, mais observée avec du bruit. On est alors dans le cadre de la statistique mathématique, en particulier, pour suivre la terminologie en vigueur, dans le champ de la régression multivariée, et plus précisément, dans la situation "p plus grand que n". Pour s'adapter aux notations usuelles dans la communauté statistique, on remplace c^* par β, Φ par X et f par y. Dans ce cas, on résoud le problème d'optimisation suivant

$$\min_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda \|b\|_{\ell_1}. \tag{4.37}$$

La solution de ce problème est souvent unique sous des conditions génériques sur la matrice Φ. La solution, notée $\hat{\beta}$, est appelée estimateur LASSO de β. La grande question est alors de savoir comment choisir λ de manière à conserver certaines propriétés du cas non-bruité, en particulier, celle qui consiste à retrouver le support de β comme étant aussi celui de $\hat{\beta}$. Une autre direction consiste à alléger la condition RIP qui est en fait facilement démontrable avec forte probabilité pour des matrices aléatoires mais NP-dur à vérifier pour des matrices présentées arbitrairement à l'utilisateur. Je vais maintenant présenter les travaux que j'ai soumis sur ces sujets.

Le Compressed Sensing et une méthode plus efficace que la minimisation ℓ_1

Dans l'article **[H]**, j'ai proposé une méthode permettant d'améliorer les performances obtenues avec la simple minimisation de la norme ℓ_1 dans le cadre du Compressed Sensing sans bruit d'observation.

Je rappelle que le problème initial était de trouver la solution la plus sparse du système linéaire $f = \Phi c$, c'est à dire de résoudre le problème (4.25). La magie de la relaxation (4.26) était de rendre retrouver la solution de (4.25) avec un simple programme linéaire. Les conditions pour réaliser un tel exploit sont par exemple la propriété de quasi-Isometrie Restreinte, ($RIP(\delta, k)$). Lorsque le nombre d'observation n'est pas suffisant, c'est à dire que

n est trop petit, la condition $(RIP(\delta, k))$ risque de ne pas être satisfaite et il faut essayer d'améliorer la stratégie au moins d'une manière algorithmique. Pour se faire, on essaie d'abord de réécrire le problème (4.25) de manière un peu différente. En particulier, (4.25) est équivalent au problème

$$\max_{z,\, c\in\mathbb{R}^p} e^t z \tag{4.38}$$

subjet aux contraintes

$$z_i c_i = 0, \quad z_i(z_i - 1) = 0 \quad i = 1, \ldots, n, \text{ et } \Phi c = f$$

où e dénote le vecteur ne contenant que des 1. Ainsi, la variable z_i joue le rôle de fonction indicatrice pour l'événement $c_i = 0$. Notons que le problème (4.-8) est clairement nonconvexe de par la présence des contraintes d'inégalité quandratiques $z_i c_i = 0, \quad i = 1, \ldots, n$. On peut alors essayer une approche par relaxation Positive Semi-Définie (SemiDefinite Programming), mais on peut facilement démontrer que l'approche naive fondée sur les constructions standard de relaxation, ne fonctionne pas. Je renvoie à l'article **[G]** pour plus de détails sur cette question. On procède alors d'une manière pragmatique en résolvant alternativement selon z et selon c.

Entrons maintenant d'un soup c con dans les détails de la méthode. Une variante de la formulation (4.-8) pourrait s'écrire de la manière suivante:

$$\max_{z\in\{0,1\}^p} e^t z \quad \text{s.t. } \|D(z)c\|_1 = 0, \quad \Phi c = f \tag{4.39}$$

où $D(z)$ est la matrice diagonale dont le vecteur diagonal est z. Si on garde les contraintes $\Phi c = f$ et $z \in \{0,1\}^p$ implicites dans (2), la fonction de Lagrange est donnée par

$$L(c, z, u) = e^t z - u\|D(z)c\|_1. \tag{4.40}$$

La fonction duale (avec valeurs dans $\mathbb{R} \cup +\infty$) est définie comme

$$\theta(u) = \max_{z\in\{0,1\}^p,\, \Phi c=f} L(c, z, u) \tag{4.41}$$

et le problème dual est alors

$$\inf_{u\in\mathbb{R}} \theta(u). \tag{4.42}$$

Evidemment, ce problème est aussi difficile que le problème original car la fonction duale est difficile à calculer explicitement à cause de la non-convexité de la fonction de Lagrange L.

Lorsque l'on restreint z à la valeur $z = e$, i.e. au vecteur dont toutes les composantes sont égales à 1, résoudre le problème

$$x_(u) = \ \text{argmax}_{z=e,\, x\in\mathbb{R}^n,\, \Phi c=f} L(c, z, u) \tag{4.43}$$

redonne exactement la solution de la relaxation ℓ_1 (4.26). On peut donc espérer légitimement qu'en optimisant la fonction en la variable z, on pourra faire mieux qu'avec la relaxation ℓ_1. L'algorithme que j'ai proposé est alors tout simplement un algorithme d'optimisation alternée en les variables c et z.

Les performances de la méthode proposée sont évaluées par simulation de type Monte Carlo sur la base de problèmes générés aléatoirement. Plus précisément, la matrice Φ est tirée au hasard avec des composantes indépendantes gaussiennes, un vecteur c^* s-sparse est tiré au hasard en choisissant un support de cardinal s uniformément parmi tous les

Algorithm 1 Alternating l_1 algorithm (Alt-l_1)

Require: $u > 0$ and $L \in \mathbb{N}_*$
  $z_u^{(0)} = e$
  $c_u^{(0)} \in \operatorname{argmax}_{c \in \mathbb{R}^p,\, \Phi c = f} L(c, z^{(0)}, u)$
  $l = 1$
  while $l \leqslant L$ **do**
    $z_u^{(l)} \in \operatorname{argmax}_{z \in \{0,1\}^p} L(c_u^{(l)}, z, u)$
    $c_u^{(l)} \in \operatorname{argmax}_{c \in \mathbb{R}^p,\, \Phi c = f} L(c, z_u^{(l)}, u)$
    $l \leftarrow l + 1$
  end while
  Output $z_u^{(L)}$ and $x_u^{(L)}$.

sous ensembles de taille s de $\{1, \ldots, p\}$ et, conditionnellement à ce support, les coordonées non nulles de c^* sont tirées suivant une loi gaussienne. On étudie alors la proportion des problèmes pour lesquels la méthode à retrouvé parfaitement le vecteur c^* en fonction de s. On obtient alors la Figure 4.

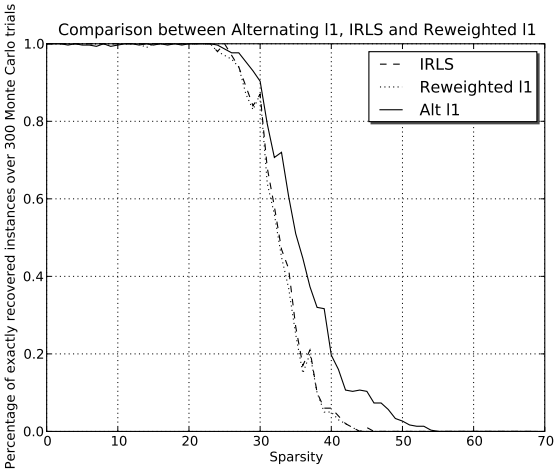

Figure .2: Taux de succès parmi 100 expériences de type Monte Carlo pour $p = 256$, $n = 100$, $L = 4$. Les componsantes non nulles de c^* ont été tirées aléatoirement selon la loi gaussienne $\mathcal{N}(0, 4)$. Les entrées de Φ ont été tirées de manière indépendantes et identiquement distribuées selon la loi gaussienne $\mathcal{N}(0, 1)$, puis normalisées à la valeur 2 pour la norme euclidienne.

Les résultats obtenus sont bien meilleurs que ceux donnés par la relaxation ℓ_1 du problème mais aussi, sont meilleurs que la relaxation Reweighted ℓ_1 de Candès, Wakin et Boyd, Enhancing sparsity by reweighted l1 minimization. J. Fourier Anal. Appl., (2008) 14 877-905. L'analyse théorique de cette méthode reste par contre encore hors de portée. Depuis, d'autres méthodes ont été publiées et il faudrait comparer leurs performances avec la méthode d'optimisation alternée proposée ici pour être sûr qu'il est encore pertinent de s'acharner à prouver des résultats théoriques à son propos. L'article est susceptible de garder un intérêt malgré la constante évoltion des approches pour la reconstruction des problèmes

sparses car il contient une analyse de la relaxation SDP naturelle et explique pourquoi elle ne marche pas. La méthode d'optimisation alternée est en elle même très intuitive et il ne serait pas surprenant que l'on retombe sur ce type d'idée dans le futur dans le cadre de l'amélioration de la reconstruction pour des problèmes plus délicats comme lorsqu'une hypothèse de sparsité sectrale est imposée sur une matrice C^* à reconstruire comme dans le problème de Matrix Completion, très à la mode depuis le Netflix Contest en filtrage collaboratif et ses applications à l'effrayant marketing ciblé dont nous sommes chaque jour un peu plus la proie désabusée.

Une facette intéressante de la prise en compte du bruit de mesure est le sujet de la prochaine section.

Le LASSO avec variance inconnue

Le LASSO est une méthode qui produit un estimateur de type moindre carrés pénalisés pour les problèmes de régression sparse en statistique, où le nombre p de covariables peut être particulièrement grand devant le nombre d'observations n. Voici le contexte: on suppose qu'on observer des données $y_1, \dots, y_n$ concaténées dans un vecteur $y \in \mathbb{R}^n$ et obtenu selon le modèle gaussien

$$y \quad = \quad X\beta + \varepsilon \tag{4.44}$$

où $X \in \mathbb{R}^{n \times p}$ est la matrice dite "de design", β est le vecteur de régression et ε est le buit de mesure, supposé suivre une loi gaussienne $\mathcal{N}(0, \sigma^2 I)$.

L'estimateur LASSO de β est donné par

$$\hat{\beta} \quad \in \quad \text{argmin}_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \lambda \|b\|_1 \tag{4.45}$$

où on retrouve une pénalisation de type ℓ_1 comme dans le Compressed Sensing, qui sert de supplétif à l'infâme "norme" (qui n'en n'est pas une) ℓ_0. Lorsque p est plus grand que n, évidemment, aucun espoir n'est permi de retrouver une bonne estimation de β. Par contre, lorsque β est supposé suffisamment sparse, on peut d'après les le c cons tirées du Compressed Sensing espérer pouvoir retrouver un estimateur pertinent et c'est ce qui se passe en théorie comme en pratique. Lorsque la condition RIP est vérifiable, les approches développées pour le Compressed Sensing sont facilement adaptable au cas bruité du modèle linéaire gaussien. Par contre, lorsqu'on ne sait pas que X satisfait une telle condition de type RIP, l'analyse est évidemment plus délicate. L'article de référence sur le sujet est E. J. Candès and Y. Plan. Near-ideal model selection by l1 minimization. Annals of Statistics, 37 (2009) 2145–2177. Cet article propose d'analyser l'estimateur LASSO dans le cas où la cohérence de la matrice X est faible, c'est à dire, dans le cas où les produits scalaires des colonnes sont très faibles en valeur absolue. Si on prend une matrice gaussienne dont les entrées sont i.i.d. $\mathcal{N}(0, \frac{1}{\sqrt{n}})$, on peut certifier facilement avec forte probabilité une cohérence

$$\mu(X) \quad = \quad \max_{j \neq j' = 1, \dots, p} |\langle X_j, X_{j'} \rangle| \tag{4.46}$$

de l'ordre de $1/\log(p)$ pour n aussi grand qu'une quantité de l'ordre de $\log(p)^3$. Ici, X_j dénote la j^{eme} colonne de X. On voit qu'on a donc de la marge et avoir des matrices très larges ayant en même temps une cohérence très faible et tendant même vers zéro. L'avantage de la cohérence est qu'elle est très rapidement calculable, alors que la constante RIP nécessite de calculer les valeurs singulières de toutes les restrictions X_T de X indexées par les ensembles $T \subset \{1, \dots, n\}$ de cardinal s et conduit à une énumération d'ordre exponentiel.

Le théorème de Candès et Plan qui nous intéresse ici est celui concernant la reconstruction exacte du support de β. Il s'agit de connaître la probabilité que $\hat{\beta}$ ait le même support que β. Dans le cas où la cohérence est de l'ordre de $1/\log(p)$ et les colonnes de X sont normalisées, c'est à dire que leur norme euclidienne est égale à 1, le Théorème de Candès et Plan donne que le support de $\hat{\beta}$ est égal à celui de β avec une probabilité de l'ordre de $1/p^2$ lorsqu'on suppose que le support de β est tiré au hasard avec une loi uniforme sur tous les sous-ensembles de $\{1,\dots,n\}$ de taille s et que le paramètre λ est de l'ordre de $\sigma\sqrt{\log(p)}$.

Dans notre article **[L]** avec Sébastien Darses, nous nous intéressons au cas où la variance σ^2 du bruit ε est inconnue. Cette situation est en fait on ne peut plus réaliste vu que les cas où elle peut être considérée comme connue sont rares en statistique appliquée. La variance est considérée connue dans le cas où X symbolise un appareil de mesure et le bruit de mesure est spécifié par le constructeur. En statistiques appliquées, on a affaire à des situations beaucoup plus variées et connaître cette variance est souvent irréaliste.

Nous avons proposé deux stratégies différentes (A) et (B) et comparé leurs performances respectives. Pour chaque stratégie, on se donne une fa c con de choisir le paramètre λ et on compare ce choix avec le cas où la variance est connue (et on se trouve dans le cas standard de l'utilisation du LASSO). Plus précisément, on pourra considérer que le choix de λ est en fait un estimateur $\hat{\lambda}$ étant donné qu'il sera une fonction des données. La méthode d'analyse consistera à exhiber un oracle $\tilde{\beta}$, qui n'est pas exactement $\hat{\beta}$, auquel on associera un oracle $\tilde{\lambda}$, qui ne sera pas exactement $\hat{\beta}$ mais qui servent de pierre angulaire à l'étude du comportement de l'estimateur. En particulier, ces oracles seront

(a) plus facilement calculable que l'estimateur $\hat{\beta}$

(b) démontrés comme égaux à $\hat{\beta}$ et $\hat{\lambda}$ respectivement, avec forte probabilité.

La Stratégie (A) est définie dans le tableau suivant

Variance connue	Variance inconnue : Strategy (A)
$\hat{\beta} \in \operatorname{argmin}_{b\in\mathbb{R}^p} \frac{\|y-Xb\|_2^2}{2} + \lambda\|b\|_1$	$\hat{\beta}_\lambda \in \operatorname{argmin}_{b\in\mathbb{R}^p} \frac{\|y-Xb\|_2^2}{2} + \lambda\|b\|_1$
$\lambda =$ cst $\sigma\sqrt{\log p}$	Choisir λ a la valeur $\hat{\lambda}$ t.q. $\hat{\lambda} = C_{\text{var}}\lambda\sigma\sqrt{\log p}$ avec: $\lambda\sigma^2 = \frac{\|y-X\hat{\beta}_{\hat{\lambda}}\|_2^2}{n}$
Probleme convexe	Probleme non-convexe
Oracle $\tilde{\beta}$	Oracle $(\tilde{\beta}, \tilde{\lambda})$
Conditions ayant lieu avec forte probabilite	Conditions similaires

La stratégie (B) est définie dans le tableau suivant

Variance connue	Variance inconnue: Strategie (B)
$\hat{\beta} \in \operatorname{argmin}_{b\in\mathbb{R}^p} \frac{\|y-Xb\|_2^2}{2} + \lambda\|b\|_1$	$\hat{\beta}_\lambda \in \operatorname{argmin}_{b\in\mathbb{R}^p} \frac{\|y-Xb\|_2^2}{2} + \lambda \|b\|_1$
$\lambda = \text{ cst } \sigma\sqrt{\log p}$	Choisir λ a la valeur $\hat{\lambda}$ t.q. : $\hat{\lambda}\|\hat{\beta}_{\hat{\lambda}}\|_1 = C \|y - X\hat{\beta}_{\hat{\lambda}}\|_2^2$
Problème convexe	probleme non-convexe
Oracle $\tilde{\beta}$	Oracle $(\tilde{\beta}, \tilde{\lambda})$
Conditions ayant lieu avec forte probabilite	Conditions similaires + Borne superieure sur $\|\beta\|_1$

Le théorème d'identification correcte du support pour la stratégie (A) est donné dans l'article. Il dépend d'une liste d'hypothèses imitant de manière assez fidèle celles de l'article de Candès et Plan. L'énumération de ces hypothèses étant fastidieuse et indigeste, on donne ici une version simplement esquissée.

Theorem 4.1 *(Esquisse) Soit $\alpha > 0$. Supposons que le support de β soit tiré au hasard uniformément parmi les sous ensembles de taille s de $\{1, \ldots, n\}$ et que*

$$s \leq s_0 := \frac{p}{\log p} \frac{C_{spar}}{\|X\|^2} \tag{4.47}$$

$$n \geq s\,(C_\circ(\alpha)\log p + 1)\,. \tag{4.48}$$

Alors, la probabilité que $\hat{\beta}$ défini par la Strategie (A) avec

$$C_{var} \in \left[\frac{(1-r)^2}{4(1+r)C_{spar}} \frac{n}{p}\|X\|^2;\ \frac{(1-r)^2}{2(1+r)C_{spar}} \frac{n}{p}\|X\|^2\right], \tag{4.49}$$

retrouve exactement le support et la configuration de signes β avec probabilité supérieure à $1 - 228/p^\alpha$.

On obtient aussi un théorème similaire pour la stratégie (B). Nous renvoyons le lecteur à l'article pour les détails.

Ces théorèmes sont fondé sur une étude fine des valeurs singulières de la matrices aléatoire X_T où T est le support aléatoire de β et X est condidérée comme déterministe. Un résultat similaire est montré dans l'article de Candès Plan, fondé sur un résultat antérieur de Tropp, lui même héritié des études menées récemment sur les matrices aléatoires de tailles finie et les travaux de Bourgain et Tzafriri. Nos résultats sont meilleurs du point de vue des constantes, mais pas des ordres de grandeurs. Nous nous approchons par la même occasion de grandeurs de l'ordre de celles qui peuvent être effectivement rencontrées en pratique, contrairement à l'article de Candès et Plan. Nous détaillerons cette sous partie, qui a donné lieu à une publication indépendante, dans la partie suivante.

Une étude expérimentale est menée afin de voir si le support est correctement retrouvé en pratique. Une des figures types parmi celles présentées dans l'article est la Figure 4. Elle

montre les histogrammes du nombre de composantes retrouvées par $\hat{\beta}$ dans la stratégie (B) (à gauche), ainsi que les histogrammes du nombre de composantes détectées à tort par $\hat{\beta}$ (à droite) avec les valeurs $C = .01, .1, .5$.

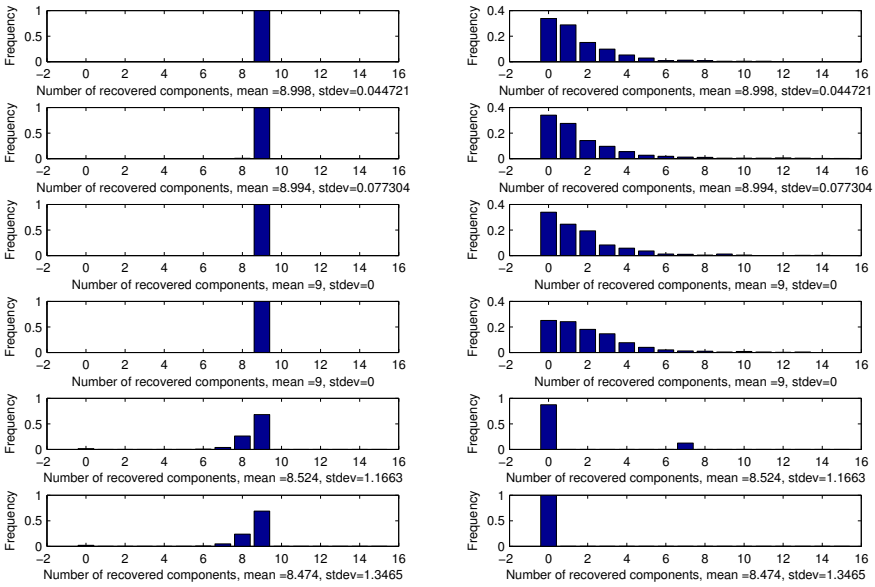

Figure .3: Histogramme du nombre de composantes retrouvées exactement par l'estimateur $\hat{\beta}$ (à gauche) et le nombre de composantes détectées à tort (à droite) pour la Strategie (B) et $C = .01, .1, .5$ avec $B = 5$. dans le cas ou les composantes non nulles de β sont tirées selon une loi gaussienne specifiee par $\beta_j = 2\mu_j + \nu_j$, ou les μ_i sont des Bernoulli ± 1 independantes et ν_j des gaussiennes standards independantes.

Il reste à donner quelques commentaires sur ces stratégies. La stratégie (B) a été la première que nous avons étudiée et la démontrastion du théorème associé est la plus difficile. Par contre, une fois étudiée, la stratégie (B) nous a permi de comprendre comment mettre en place la stratégie (A) qui semble en fait plus naturelle a posteriori. Curieusement, ces deux stratégies n'ont pas le même comportement en pratique: la stratégie (A) fonctionne en gros comme le LASSO avec variance connue. Elle permet de plus d'estimer la variance assez correctement (voir les simulations dans l'article) dans les mêmes configurations d'expérience que le LASSO. La stratégie (B) en revanche marche beaucoup mieux dans le cas où le rapport signal sur bruit est faible (les coefficients de β ne sont pas très grands devant l'écart-type σ). Ceci est en accord avec le fait que dans la stratégie (B), on demande de plus qu'une borne supérieure sur la norme ℓ_1 de β soit satisfaite. Nous pensons que la stratégie (B) devrait être étudiée plus en détail en mettant en valeur ce faible rapport signal sur bruit mais nous n'avons pas encore trouvé les outils techniques de probabilité permettant d'approchondir les résultats dans ce sens.

Quasi-Isométrie Restreinte et incohérence

Le problème de retrouver le support d'un vecteur de régression par le LASSO dans le cas de la variance connue comme dans le cas où la variance est inconnue demande d'étudier les valeurs singulières des matrices aléatoires X_T où T est un sous ensemble d'indices de $\{1,\dots,n\}$ tiré uniformément sur tous les sous-ensembles d'indices de taille s. Ce problème a été étudié dans l'article de Candès et Plan cité dans la partie précédante a partir de résultats tiré de l'article J. Tropp, Norms of random submatrices and sparse approximation, C. R. Acad. Sci. Paris, Ser. I, Vol. 346, (2008), pp. 1271-1274. Nous nous sommes penché sur l'analyse de ce problème avec Sébastien Darses, dans le cadre de notre étude sur le LASSO avec variance inconnue, décrit dans la partie précédente, car nous avons cru possible d'améliorer les résultats de Tropp et Candès-Plan en utilisant de nouvelles inégalités de déviation matricielles récemment mise en valeur dans l'article de J. Tropp, User-friendly tail bounds for sums of random matrices, Found. Comput. Math., Vol. 12, (2012) num. 4, pp. 389–434. L'article [**K**] donne des résultats sur la probabilité qu'une sous matrices X_T tirée au hasard selon la loi susdécrite soit bien conditionnée qui sont les meilleurs jusqu'à maintenant. Ils ont déjà été cités et utilisés comme brique intermédiaire dans la littérature relative au Compressed Sensing et aux algorithmes pour les problèmes inverses sous contrainte de sparsité.

Commen c cons par rappeler des résultats standards sur les sommes de variables aléatoires indépendantes. Le théorème de concentration de Bernstein est bien connu et très utilisé dans la littérature probabiliste. Il s'énonce comme suit.

Theorem 4.2 *Soient $X_1,\dots,X_n$ des variables aléatoires réelles indépendantes et centrées et telles que $|X_i| \leq M$ presque sûrement. Alors, on a*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq u\right) \quad \leq \quad \exp\left(-\frac{\frac{1}{2}u^2}{\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right]+\frac{1}{3}M\ u}\right) \tag{4.50}$$

Le théorème de Hoeffding est un autre théorème de déviation qui ne tient pas compte de la variance des termes mais simplement de la borne supérieure sur la valeur absolue de chaque terme. Il s'énonce comme suit.

Theorem 4.3 *Soient $X_1,\dots,X_n$ des variables aléatoires réelles indépendantes telles que pour tout $i=1,\dots,n$, $a_i \leq X_i \leq b_i$ presque sûrement. Alors, on a*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]+u\right) \quad \leq \quad \exp\left(-\frac{2u^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right) \tag{4.51}$$

La preuve de ces théorèmes suit une procédure assez similaire basée sur l'inégalité de Chernov. Posons $Y_i = X_i - \mathbb{E}\left[X_i\right]$. On a

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq u\right) \quad = \quad \mathbb{P}\left(\exp\left(\lambda\ \sum_{i=1}^{n} Y_i\right) \geq \exp\left(\lambda\ u\right)\right) \tag{4.52}$$

$$\leq \quad \frac{\mathbb{E}\left[\exp\left(\sum_{i=1}^{n}\lambda Y_i\right)\right]}{\exp\left(\lambda\ u\right)}. \tag{4.53}$$

Par indépendance, on peut ensuite développer et obtenir

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq u\right) \quad \leq \quad \frac{\mathbb{E}\left[\prod_{i=1}^{n}\exp\left(\lambda Y_i\right)\right]}{\exp\left(\lambda\ u\right)}, \tag{4.54}$$

puis en utilisant l'indépendance, on obtient,

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq u\right) \quad \leq \quad \frac{\prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\lambda Y_i\right)\right]}{\exp\left(\lambda\ u\right)}, \tag{4.55}$$

et il suffit alors d'avoir une estimation de $\mathbb{E}\left[\exp\left(\lambda Y_i\right)\right]$ puis d'optimiser en λ pour obtenir l'inégalité prévue. Selon les hypothèses, on obtient des bornes sur $\mathbb{E}\left[\exp\left(\lambda Y_i\right)\right]$ qui ont des formes un peu différentes selon les hypothèses. Si Y_i est simplement considérée comme sous-gaussienne, alors

$$\mathbb{E}\left[\exp\left(\lambda Y_i\right)\right] \quad \leq \quad e^{c^2\lambda^2}. \tag{4.56}$$

pour une constante à déterminer et l'optimisation en λ est triviale.

Dans notre cas d'étude, nous avons à traiter une somme de matrices de rang 1 pondérée par des variables de Bernoulli. On peut utiliser une approche identique à celle utilisée pour les déviations de sommes de variables aléatoires mais il se passe un couic. En effet, pour des matrices qui ne commutent pas, nous d'avons pas $\exp(A+B) = \exp(A)\exp(B)$ et l'approche tombe a l'eau. Il est malgré tout possible de contourner le problème grâce à une inégalité de convexité matricielle due à Lieb. Elle dit la chose suivante.

Theorem 4.4 *Soit H une matrice symétrique réelle positive semi-définie. La fonction $X \mapsto$ trace$\exp\left(H + \log(X)\right)$ est concave sur le cône des matrices symétriques réelles positives semi-définies.*

A partir de cette inégalité très puissante, on peut en déduire une borne très générale en suivant la même méthode que pour le cas réel. On procède de la manière suivante: on a d'abord pour toute matrice symétrique réelle aléatoire Y

$$\begin{aligned}
\mathbb{P}\left(\lambda_{\max}(Y) \geq u\right) &= \mathbb{P}\left(\exp\left(\theta\ \lambda_{\max}(Y) \geq \exp(\theta u)\right)\right) && (4.57)\\
&\leq \mathbb{P}\left(\exp\left(\theta\ \ \mathrm{trace}(Y) \geq \exp\left(\theta\ u\right)\right)\right) && (4.58)\\
&\leq \frac{\mathbb{E}\left[\exp\left(\theta\ \ \mathrm{trace}(Y)\right]\right)}{\exp\left(\theta\ u\right)} && (4.59)
\end{aligned}$$

puis on pose $Y = \sum_{j=1}^{p} Z_j$ et on utilise l'inégalité de Lieb récursivement en conditionnant par rapport aux termes successifs de la somme. On obtient alors, en optimisant par rapport à θ,

Theorem 4.5 *Soient $Z_1, \ldots, Z_p$ des matrices symmétriques réelles indépendantes. Alors,*

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{j=1}^{p} Z_j\right) \geq u\right) \quad \leq \quad \inf_{\theta>0} \frac{\mathrm{trace}\exp\left(\log\left(\mathbb{E}\left[e^{\theta Z_j}\right]\right)\right)}{e^{\theta u}} \tag{4.60}$$

Le reste de l'étude revient à borner convenablement $\mathbb{E}\left[e^{\theta Z_j}\right]$ en fonction de la cohérence $\mu(X)$. C'est à nouveau un peu technique mais nous renvoyons le lecteur à l'article pour les détails. On aboutit alors à notre théorème principal.

Theorem 4.6 *Soit $r \in (0,1)$, $\alpha \geq 1$. Supposons*

$$\begin{aligned}
\mu(X) &\leq \frac{r}{(1+\alpha)\log p} && (4.61)\\
s &\leq \frac{r^2}{(1+\alpha)e^2}\ \frac{p}{\|X\|^2 \log p}. && (4.62)
\end{aligned}$$

Alors, on a la borne suivante:

$$\mathbb{P}\left(\|X_T^t X_T - I\| \geq r\right) \quad \leq \quad \frac{1944}{p^\alpha}. \tag{4.63}$$

Modèles de mélanges pour les matrices de design et performance en prédiction

Après l'étude du problème de retrouver le support d'un vecteur de régression sparse sous des contraintes d'incohérence, c'est à dire en supposant que la matrice X avait des colonnes "suffisamment" orthogonales, je me suis tourné vers le problème de savoir comment traiter le cas où la cohérence $\mu(X)$ de la matrice X ne satisfaisait pas les conditions habituelles de petitesse. Dans ce cas, il semble difficile d'imaginer que le support de β puisse être effectivement retrouvé avec forte probabilité. Malgré cette fatalité insurmontable, il existe des espoirs certains sur les possibilité du LASSO à produire un estimateur raisonnable. Pour comprendre le potentiel de cet estimateur dans des situations moins favorables, il suffit de se tourner vers un autre critère de performance: l'erreur quadratique commise en prédiction.

L'objectif de l'article **[O]** est de borner l'erreur quadratique pour un modèle très particulier de matrices de design où l'hypothèse de petite cohérence, comme dans le travail de Candès et Plan sur le LASSO, n'est pas respectée. Afin de pouvoir contrôler de manière la plus précise l'erreur de prédiction, on va avoir recours à un modèle pour la génération de la matrice X. Le modèle que j'ai proposé est le suivant: on suppose qu'on a de manière sous-jascente une matrice de design $\mathfrak{C}$ dans $\mathbb{R}^{n\times K}$ qui, elle verife la condition de petite cohérence et que les colonnes de X des perturbations de ces colonnes, avec nombreuses éventuelles répétitions permises. On obtient donc une sorte de modèle de mélange pour les colonnes de X, les centres desquels sont les colonnes de $\mathfrak{C}$ et K le nombre de composantes du modèle de mélange. Avec ce type de modèle, il faut évidemment s'attendre à ce que l'estimateur LASSO se "trompe" de colonnes, i.e. de covariables. En effet, deux colonnes de X obtenues par perturbation d'une même colonne de X_o pourront aisément être confondues dans leurs effets sur y. L'idée à garder est que ce phénomène n'est pas très grave dans le modèle proposé car choisir l'un ou l'autre entre les avatars d'une même colonne $X_{j,o}$ de X_o est tout à fait supportable car cela revient avec l'une ou l'autre à désigner cette colonne $X_{j,o}$ comme vraie responsable des effets sur y. Un tel modèle paraît réaliste si l'on n'a pas d'objectif de recherche de causalité, ce qui est bien le cas pour un modèle de régression multivariée comme celui traité par les méthodes de type LASSO.

La matrice de design X va être supposée choisie de la manière suivante. Soit $\mathcal{K}$ un sous ensemble aléatoire de $\{1,\ldots,K\}$ ayant pour cardinal s^*, supposé tiré avec la loi uniforme sur tous les sous-ensembles de $\{1,\ldots,K\}$ de même cardinal. On suppose ensuite que, conditionnellement à $\mathcal{K}$, chaque colonne de X_o est tirée selon une loi de mélange gaussien n-dimensionnel dont chaque composante est centrée sur une des colonnes de $\mathcal{K}$, i.e.

$$\Phi(x) \quad = \quad \sum_{k\in\mathcal{K}} \pi_k \phi_k(x), \tag{4.64}$$

où

$$\phi_k(x) \quad = \quad \frac{1}{(2\pi\mathfrak{s}^2)^{\frac{n}{2}}} \exp\left(-\frac{\|x-\mathfrak{C}_k\|_2^2}{2\mathfrak{s}^2}\right), \tag{4.65}$$

et $\pi_k \geq 0$, $k\in\mathcal{K}$ et $\sum_{k\in\mathcal{K}}\pi_k = 1$. On va dénoter par n_k le nombre aléatoire de colonnes de X_o qui ont été tirée suivant la distribution $\mathcal{N}(\mathfrak{C}_k, \mathfrak{s}^2 I)$, $k = 1,\ldots,K_o$. Ainsi, on a

obligatoirement $\sum_{k\in\mathcal{K}} n_k = p$. Après avoir obtenu la matrice X_o, la matrice X est dérivée de X par normalisation euclidienne des colonnes, i.e. $X_j = X_{o,j}/\|X_{o,j}\|_2$ pour chaque j dans $\{1,\dots,p\}$.

Le résultat principal de l'article **[O]** est le théorème dont on peut donner l'esquisse suivante. Les détails des hypothèses sont à regarder dans l'article.

Theorem 4.7 *Soit $\alpha \in (0,1)$. Posons $\lambda = 2\sigma\sqrt{2\alpha\ \log(p)}$. Alors, avec probabilité plus grande que $1-p^{-\alpha}$,*

$$\frac{1}{2}\|Xh\|_2^2 \quad \leq \quad s^*\frac{3}{2}r_*\ \lambda\left(\frac{3}{2}\ \lambda + \sqrt{1+r_*}\ \delta\|\mathfrak{C}_T\beta_T\|_2\right) + \frac{1}{2}\delta^2\|X\beta\|_2^2 \tag{4.66}$$

avec $r_* = 1.1\cdot r\ \ (1.1 + 0.11\cdot r)$ et pour n'importe quel δ tel que

$$\begin{aligned}
\delta \quad \geq \quad & 4\mathfrak{s}\ \left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right)\left(1 + 8\sqrt{2}\ \sqrt{\alpha\log(p) + \log(2n+2)}\sqrt{s^*\rho_{\mathfrak{C}}}\right) \\
& + \left(12\ C_{\int}\ \mathfrak{s}\ \sqrt{n}\ r_{\max} + \alpha\log(p)\ \mu_{\max}\right)\ \sqrt{s^*\rho_{\mathfrak{C}}} \\
& + 4\mathfrak{s}\sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\left(1 + 2\sqrt{2}\ \sqrt{\rho_{\mathfrak{C}}}\sqrt{\alpha\log(p) + \log(2n+2))}\right) \\
& + \left(24\ r^*_{\max}\ \mathfrak{s}\sqrt{\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\ C^*_{\int}} + \mu^*_{\max}\ \alpha\log(p)\right)\ \sqrt{\rho_{\mathfrak{C}}},
\end{aligned} \tag{4.67}$$

où les diverses constantes et leur contraintes mutuelles sont spécifiées dans les hypothèses de l'article.

L'énoncé du théorème ainsi que les hypothèses sont clairement très indigestes. Les hypothèses ne demandent en substance que des choses très naturelles: il faut que les colonnes de $\mathfrak{C}$ soient très orthogonales, comme dans l'article de Candès et Plan sur le LASSO, et que les colonnes de X_o soient tirées avec un variance très petites autour de leurs centres respectifs, de manière que les clusters soient suffisamment séparés afin de ne créer une quelconque confusion entre les centres eux mêmes qu'avec une probabilité très petite, i.e. de l'ordre de $p^{-\alpha}$.

Une astuce pour les matrices très mal conditionnées: accoler une matrice gaussienne

Le but du dernier travail que je présenterai ici concerne aussi, comme dans la partie précédente, le cas du modèle de régression sparse avec grand nombre de covariables, i.e. $p \gg n$. La stratégie est différente: on ne suppose plus de modèle sous-jascent comme par exemple un modèle de mélange gaussien dans le travail précédent. On va procéder en utilisant une astuce algorithmique: on accole à la matrice de design pré-existante dont les colonnes seront supposées normalisées, une matrice gaussienne aux composantes i.i.d. $\mathcal{N}(0, \frac{1}{\sqrt{n}})$, normalisée pour la norme euclidienne. C'est en examinant les conditions d'optimalité pour le LASSO qu'on peut se rendre compte qu'une telle astuce est possible et que l'estimateur du LASSO jouit de propriétés équivalentes à celles du cas où la cohérence est, par design petite. En effet, quel que soit le support de $\hat{\beta}$, on peut trouver dans la matrice gaussienne postnormalisée qu'on a accolé une sous matrice de même taille et qui soit, elle, très bien conditionnée avec forte probabilité. Il faut pour se faire choisir d'accoler une matrice gaussienne de taille

relativement grande. Le calcul de la taille minimal prend un certain effort et tient une bonne place dans la démonstation du théorème principal de l'article. Sans entrer dans plus de détail, on parvient à démontrer que l'estimateur LASSO avec une matrice gaussienne postnormalisé accolée peut commettre une erreur de prédiction du même ordre de grandeur que si la matrice de design elle même avait initialement eu la propriété de petite cohérence.

Pour faciliter l'étude, on peut introduire une quantité intermédiaire pour caractériser la qualité de la matrice de design.

Definition 4.8 *L'indice $\gamma_{s,\rho_-}(X)$ associé à la matrice X est défini par*

$$\gamma_{s,\rho_-}(X) \quad = \quad \sup_{v\in B(0,1)} \inf_{I\subset \mathcal{S}_{s,\rho_-}} \|X_I^t v\|_\infty. \tag{4.68}$$

On note facilement que la fonction $s \mapsto \gamma_{s,\rho_-}(X)$ est croissante. Un fait important à noter est que la fonction $X \mapsto \gamma_{s,\rho_-}(X)$ est décroissante au sens que si $X' = [X, x]$ où x est un vecteur colonne de $\mathbb{R}^n$, alors $\gamma_{s,\rho_-}(X) \geq \gamma_{s,\rho_-}(X')$.

Un des intérêt de cette quantité est que l'on peut démontrer que pour n et s constants, la quantité $\gamma_{s,\rho_-}(X)$ est très petite pour p suffisamment grand, au moins pour des matrices aléatoires telles que les gaussiennes post-normalisées. La propriété de monotonie permet alors de dire qu'en concaténant une matrice gaussienne à une matrice de design donnée, la matrice qui en résulte à un indice γ_{s,ρ_-} au moins aussi bon que celui de la matrice gaussienne post-normalisée accolée.

Le théorème principal de l'article **[A14]** écrit sur ces travaux est esquissé dans l'énoncé suivant.

Theorem 4.9 *Soit $\rho_- \in (0,1)$. Soit ν tel que*

$$\nu\ \gamma_{\nu n,\rho_-}(X) \leq \frac{\rho_-^2}{n\ \rho_+}. \tag{4.69}$$

Suposons que $s \leq \nu n$. Supposons que β ait pour support $S \in \mathcal{S}_{s,\rho_-^+}(X)$ et que

$$\lambda \quad \geq \quad \sigma\left(B_{X,\nu,\rho_-,\rho_+}\sqrt{2\alpha\ \log(p)+\log(2\nu n)} + \sqrt{(2\alpha+1)\ \log(p)+\log(2)}\right) \tag{4.70}$$

avec

$$B_{X,\nu,\rho_-,\rho_+} \quad = \quad \frac{\nu n\ \gamma_{\nu n,\rho_-}(X)\ \rho_+}{\rho_-^2 - \nu n\ \gamma_{\nu n,\rho_-}(X)\ \rho_+}. \tag{4.71}$$

Alors, avec probabilité plus grande que $1-p^{-\alpha}$, on a

$$\frac{1}{2}\|X(\hat\beta-\beta)\|_2^2 \quad \leq \quad s\ C_{n,p,\rho_-,\alpha,\nu,\lambda} \tag{4.72}$$

avec

$$C_{n,p,\rho_-,\alpha,\nu,\lambda} \quad = \quad \frac{\lambda+\sigma\sqrt{(2\alpha+1)\ \log(p)+\log(2)}}{\rho_-^2}\left(\sigma\sqrt{2\alpha\ \log(p)+\log(2\nu n)}+\lambda\right) \tag{4.73}$$

Ces résultats ont été soumis dans **[N]** et attend son verdict ...

5 Autres travaux

Outre les thématiques que j'ai présentes plus haut, j'ai aussi collaboré dans des directions variées sur des problèmes pratiques en informatique et en ingénierie.

number of processors	1	2	3	4	5	**6**
time	60	30	20	20	20	**10**

Table .1: Example of allocation for a job which originally required 6 processors

Classification d'images pour l'analyse de l'avancement des glaciers au Spitzberg

J'ai été contacté par Jean-Michel Friedt, travaillant pour la société Senseor, afin d'analyser des stocks très grand d'images prises à intervalles fixes par des caméras placées le long d'un glacier au Spitzberg. Ce problème paraissa vraiment très simple de prime abord mais il s'est avéré très délicat. J'ai voulu appliquer des méthodes de classifications pour différencier les saisons (luminosité, jeux d'ombres) , les conditions météo (pluie, neige, beau temps), les photos sans intérêt (brume, objectif mouillé) ... Or à ma grande suprise, les algorithmes classiques ont eu beaucoup de mal à détecter et mettre dans des classes a part les photos sans intérêt ... sans parler de faire quoique ce soit de plus pécis !

Après de multiples essais, une méthode basée sur l'approximation de la Tranformée de Fourier Discrète bi-dimensionnelle par des matrices de rang faibles recodée en vecteurs puis injectés dans une méthode de spectral clustering a pu donner des résultats enfin satisfaisants. Ils ont été publiés dans l'article de conférence **[C2]**.

Scheduling pour des clusters

Ce travail a commencé à l'initiatice de Jean-Marc Nicod et Laurent Philippe et Lamiel Toch du département d'informatique à l'université de Franche-Comté. Il s'agissait de répartir des tâches sur des processeurs de manière à obtenir le make-span le plus petit possible. Le make-span est le temps que vont prendre les processeurs pour accomplir toutes les tâches. On se place ici dans une optique statique, où les jobs (tâches) sont définies à l'avance et aucune autre tâche n'arrive dans l'intervalle qui nécésssiterait de modifier la plannification. Nous avons proposé une nouvelle représentation du problème pour mettre en valeur les propriétés de sparsité de la solution et utiliser des techniques très reliées aux outils utilisés en Compressed Sensing. Ces résultats ont été publiés dans **[P]**.

Entrons maintenant un peu plus dans les détails. On considère un ensemble de jobs $J = \{J_i, 1 \leqslant i \leqslant n\}$. Les jobs sont des tâches apriori réalisables sur plusieurs processeurs en parallèle. Les variables permettant de contrôler le traitement de chaque job sont les suivantes: son temps d'execution et le nombre de processeurs qui lui sont aloués. Un job rigide ne peut pas être exécuté sur un nombre différent de processeurs que celui demandé lors de la requête. Un job est dit **moldable** s'il peut être excécuté sur un nombre flexible de processeurs que l'on peut choisir comme variable d'optimisation. On prendra plus précisément le modèle de DUTOT (Scheduling moldable BSP tasks Pierre-Fran c cois Dutot, Alfredo Goldman, Fabio Kon, Marco Netto, 11th Workshop on Job Scheduling Strategies for Parallel Processing 3834 (2005) 157–172). Soit $reqtime_i$ le temsp d'exécution du job J_i qui requiet au plus $reqproc_i$ processeurs. Soit $t_i(n)$ le temps d'exécution du job J_i si n processeurs lui sont alloués. La relation entre $t_i(n)$ et n est :

$$\forall i,\ \forall n \leqslant reqproc_i,\ t_i(n) = \left\lceil \frac{reqproc_i}{n} \right\rceil reqtime_i$$

La table .1 donne un exemple avec $reqtime_i = 10$ unités de temps et $reqproc_i = 6$ processeurs. Dans la suite, on suppose que tous les jobs sont moldables. On peut alors se représenter un jobs comme un rectangle dont la base est le nombre d'unités de temps

utilisés et la hauteur le nombre de processeurs utilisés. La surface du rectangle représente alors la contrainte réelle de réalisation du job, c'est à dire le temps total processeur dont il a besoin pour être effectué. Le but est dont de placer des rectangle qui sont déformables dans un grand rectangle dont la hateur est une contrainte égale au nombre total de processeurs disponibles.

Estimation de fréquence pour un RADAR

En 2012, j'ai été recontacté par Jean-Michel Friedt de la société Senseor pour un problème d'estimation de précise de fréquences dans un appareillage RADAR. Ce problème apparaît de fa c con récurrente en traitement du signal mais des solutions satisfaisantes viennent seulement d'être proposées. L'idée est simple: un théorème de Shannon dit que si on a un signal dont le support fréquentiel est compact, alors, il suffit d'échantillonner 2 fois plus vite pour ne perdre aucune information. Le Compressed Sensing s'est notamment fait beaucoup de publicité en annon c cant que ce théorème pouvait être largement contourné car la plupart du temps, les signaux ont une tranformée de Fourier très peut étendue et jouissent de structures supplémentaires qui permettent de reconstruire des signaux à partir d'un nombre extrêment moindre d'échantillon que le nombre prévu par le théorème de Shannnon. Malheureusement, ce qu'on appelle échantillonnage en Compressed Sensing reviens souvent à changer de base pour prendre les mesures, en choisissant une base incohérente avec la base dans laquelle le signal est sparse, et à prendre quelques coefficients dans la nouvelle base. Pour le problème que nous avons ici, cela ne servirait malheureusement à rien de changer de base: nous disposons d'échantillons temporels régulièrement espacés du signal à étudier. Par contre, pour le problème que m'a proposé J.-M. Friedt dans cette étude, une information importante est à prendre en compte: nous savons a priori que le signal n'est composé que d'une somme d'un très petit nombre de termes sinusoidaux possiblement amortis. Même si la fréquence maximale est grande, nous pouvons espérer utiliser cette sparsité d'une manière efficace. Pour ce travail, il fallait transformer le problème en un problème de régression. Un mois plus tard, Ben Recht faisait une présentation sur ce type de problème de l'université de Wisconsin, Madison, USA au European Meeting of Statisticians à Istanbul et nous avons profité d'un code qu'il a accepté de nous fournir. Nous avons alors pu le modifier en suivant les résultats que nous avons publiés avec Sébastien Darses dans [**K**] sur le problème de ne pas connaître la variance a priori dans un problème de régression. Nous avons ainsi pu retrouver des fréquences avec une précision dépassant largement ce qu'il est permis d'obtenir avec la FFT (qui ne prennent pas en compte la sparsité) et des méthodes de type Pissarenko (qui la prenne en compte mais d'une manière différente) . Les résultats d'application au problème de RADAR ont été simples à fournir et un réel plaisir à transmettre, du fait de pouvoir répondre rapidement et efficacement à une question pratique. Ils ont été publiés dans un article que nous n'avons pas inclus, vue la contribution minime qu'il représente comparée à celle de mes co-auteurs travaillant dans la communauté capteurs et électronique.

6 Nouvelles thématiques: perturbations spectrale de matrices symétriques réelles

Suite aux travaux réalisés dans le cadre de l'étude du LASSO en statistique et en particulier à l'analyse des valeurs propres extrèmes de matrices du type $A_T^t A_T$ où A_T est une matrices rectangulaire aléatoire, je me suis intéressé de plus en plus à la perturbation du spectre des matrices symétriques réelles et des matrices Hermitiennes. L'objectif initial était d'attaquer le problème de contrôler la constante d'Isométrie Restreinte dans le cas de matrices A où

les lignes sont tirées aléatoirement comme dans le cas des matrices de Fourier discrète. Un problème associé très connu est la conjecture Λ_1 qui nous a été expliquée par Stefan Neuwirth du Laboratoire de Besançon. Sans espéré apporter de contribution spectaculaire à ce problème, il m'a paru passionnant de m'y plonger en collaboration avec S. Darses afin de se donner le temps de comprendre mieux les origines de ce problème en analyse harmonique et les résultats spectaculaires obtenus par J. Bourgain et beaucoup d'autres. Sans ce type d'entrée, une fois de plus, la curiosité la plus simple vis à vis des questions qui ont préoccupé de grands chercheurs serait resté inaccessible, alors qu'en consacrant un peu de temps à un sous problème formulé simplement en termes d'algèbre linéaire nous a permis de rencontrer beaucoup de belles idées dans le domaine et d'apprendre une bonne quantité de jolies mathématiques provenant de toutes époques.

Perturbations de rang un et ajout de colonnes

Nous nous sommes mis en tête de comprendre comme le spectre d'une matrice était perturbé par adjonction d'une colonne. L'idée initiale était de pouvoir appliquer une approche de type chainage qui aurait pu nous aider en passant par le fameux résultat de Talagrand sur la comparaison de l'espérance du suprémum d'un processus aux accroissements sous-gaussiens avec une inégrale impliquant des quantités issus de processus purement gaussiens. Nous n'avons pas pu mettre ce programme en oeuvre étant donnée la difficulté de la tâche mais nous nous sommes bien régalés.

Nous avons récemment soumis un article à Linear Algebra and Applications [**Q**] résumant ce que nous avons trouvé sur la perturbation des valeurs propres extrêmes. Il en sort des inégalités toutes simples qui s'avèrent meilleures ou moins contraignantes que les inégalités récentes sur le sujet comme celle de Li-Li et Nadler par exemple. L'article fait l'objet de la Section 5. Ces bornes utilisent des techniques qui ressemblent d'assez près à celles utilisées par Spielman et ses co-auteurs dans létude du problème d'inversibilité restreinte. Cela nous a aussi donné l'occasion de lire avec plus de recul leurs travaux autour de ces sujets ainsi que l'application aux matrices de covariances pour lesquels une riche litérature a récemment paru autour de A. Pajor et N. Tomczak-Jeagerman, puis finalement des r'esultats impressionnants de Koltchinski et Mendelson, résumant les résultats accumulé en un argument très simple récemment rendu encore plus pédagogique par J. Tropp.

Soit $X \in \mathbb{R}^{d\times n}$ une matrice et $x \in \mathbb{R}^d$ un vecteur colonne. Il y a deux approches pour étudier la matrice (x, X) obtenue en ajoutant la colonne x à la matrice X:

(A1) On considère la matrice

$$A \;=\; \begin{bmatrix} x^t \\ X^t \end{bmatrix} \begin{bmatrix} x & X \end{bmatrix} = \begin{bmatrix} x^t x & x^t X \\ X^t x & X^t X \end{bmatrix}; \tag{6.74}$$

(A2) On considère la matrice

$$\widetilde{A} \;=\; \begin{bmatrix} x & X \end{bmatrix} \begin{bmatrix} x^t \\ X^t \end{bmatrix} = XX^t + xx^t.$$

Dans l'approche (A1), on considère les valeurs propres de la matrice hermitienne A qui n'est rien d'autre que la matrice X^tX augmentée d'une matrice de type "tête de flèche" ("arrowhead matrix" en anglais).

D'un autre côté, l'approche (A2) considère une matrice $\widetilde{A}$ de taille $d \times d$, qui est une perturbation de rang 1 de la matrice XX^t. Les matrices A et $\widetilde{A}$ ont les mêmes valeurs

propres non-nulles. En particulier, $\lambda_{\max}(A) = \lambda_{\max}(\widetilde{A})$. De plus, les valeurs singulière de la matrice (x, X) sont les racines carrées de la matrice A.

Nous nous sommes en fait amusés à étudier plus généralement le problème de comparer les valeurs propres de

$$A \quad = \quad \begin{bmatrix} c & a^t \\ a & M \end{bmatrix}, \tag{6.75}$$

à celles de M, où $a \in \mathbb{R}^d$, $c \in \mathbb{R}$ et $M \in \mathbb{R}^{d\times d}$.

Bien que partis de considérations liées au LASSO et à l'étude des matrices de covariance et des constante d'isométrie restreintes, nous nous sommes vite rendu compte que ce type de problème apparaît dans une grande quantité de problèmes en analyse numŕique [4], [4], théorie des graphes [8], théorie du contrôle [31], statistique [28], etc ...

Perturbation des valeurs singulières: quelques résultats existants

Obtenir des estimations précises sur les valeurs propres d'une somme de deux matrices (par exemple $X+P$, avec P une perturbation) est une tâche très difficile en général. Les inégalités de Weyl et Horn, par exemple, peuvent être utilisés et les bornes qui en résultent peuvent être améliorées si la perturbation P est faible par rapport à X (voir par exemple [**?**, Chap. 6]). Les travaux de [2] et [3], pour ne nommer qu'eux, on permis de comprendre comment on pouvait obtenir des résultats plus précis avec forte probabilité lorsque la perturbation était une matrice aléatoire.

Weyl's inequalities. La référence [35] donne de multiples résultats historiques sur les valeurs propres de sommes de matrices hermtiennes ou symmétriques réelles. Les inégalité de Weyl sont les résultats suivants:

Theorem 6.1 (Weyl) *Soient B et B' deux matrices symétriques $\mathbb{R}^{d\times d}$ et soient $\lambda_j(B)$, $j = 1, \ldots, d$, (resp. $\lambda_j(B')$), les valeurs propres de B (resp. B'). Alors,*

$$\lambda_{i+j-1}(B + B') \quad \leq \quad \lambda_i(B) + \lambda_j(B'),$$

pour $i, j \geq 1$ et $i + j - 1 \leq n$.

The arrowhead perturbation. On peut alors utiliser les inégalités de Weyl pour obtenir le résultat suivant.

Proposition 6.2 *We have*

$$\lambda_1(A) \quad \leq \quad \max\{c, \lambda_1(M)\} + \|a\|_2.$$

Preuve: Les inégalités de Weyl donnent alors, en prenant $i, j = 1$,

$$\lambda_1(A) \quad \leq \quad \lambda_1\left(\begin{bmatrix} c & 0 \\ 0 & M \end{bmatrix}\right) + \lambda_1(E) \tag{6.76}$$

with

$$E = \begin{bmatrix} 0 & a^t \\ a & 0 \end{bmatrix}.$$

De plus, en passant par la formulation variationnelle de la valeur propres la plus grande, on obtient $\lambda_1(E) = \|a\|_2$. En combinant ceci avec (2.0), le résultat voulu tombe. □ Le fait le plus important à retenir de cette inégalité est que si x est orthogonal à toutes les colonnes de

X, alors $a = 0$ et la perturbation n'a pas d'impact sur la valeur propre la plus grande tant que $c \leq \lambda_1(M)$. Cette observation élémentaire peut être extrapolée à des situations plus compliquées, comme par exemple dans le modèle de spiked covariance, où une transition de phase est observée et dont dépend la détectabilité d'un signal dans du bruit, selon le niveau d'énergie du signal [28, Theorem 2.3].

The rank-one perturbation. Si nous ne voulons qu' étudier la perturbation de la plus grande valeur propre, alors nous pouvons envisager le rang d'un perturbation décrite par (A2). dans ce cas, l'inégalité de Weyl's donne le résultat suivant.

Proposition 6.3 *We have*

$$\lambda_1(A) \quad \leq \quad \lambda_1(M) + \|x\|_2^2.$$

Preuve: Using that $\lambda_1(A) = \lambda_1(\widetilde{A})$ and $\lambda_1(M) = \lambda_1(\widetilde{M})$, we obtain from Theorem 2.1 :

$$\lambda_1(A) \quad \leq \quad \lambda_1(M) + \lambda_1(xx^t).$$

Since $\lambda_1(xx^t) = \|x\|_2^2$, the conclusion follows. □ Le principal inconvénient de cette inégalité est qu'elle ne tient pas compte de la géométrie du problème et en particulier l'angle entre X et le nouveau vecteur x que nous voulons ajouter à X.

Une inégalité de Li et Li. Ils prouvent une inégalité générale concernant la perturbation des valeurs propres en vertu de perturbations bloc-diagonales. Nous précisons leur résultat [24, Theorem 2], dans notre context:

$$|\lambda_1(A) - \max(c, \lambda_1(M))| \quad \leq \quad \frac{2\|a\|^2}{\eta + \sqrt{\eta^2 + 4\|a\|^2}}, \tag{6.77}$$

avec $\eta = \min\{|c - \lambda_i(M)|, 1 \leq i \leq d\}$. dans leur article, $\widetilde{\lambda}_1$ est en fait $\max(c, \lambda_1(M))$ pour nous. On renvoit à [24] pour l'historique de ce type d'inégalité.

une inégalité due à Ipsen et Nadler. Dans [21], les auteurs proposent une borne supérieur aux valeurs propres de $\widetilde{A}$ pour le problème de perturbation de rang 1 (A2). Le résultat suivant est un corollaire de leur résultat principal.

Theorem 6.4 *Soit $\widetilde{M} \in \mathbb{C}^{d\times d}$ une matrice Hermitienne et $x \in \mathbb{C}^d$. Soit V_1 (resp. V_2) le vecteur propre associé à $\lambda_1(\widetilde{M})$ (resp. $\lambda_2(\widetilde{M})$). Soit $\widetilde{A} = \widetilde{M} + xx^t$. Alors*

$$\lambda_1(\widetilde{M}) + \delta_{\min} \quad \leq \lambda_1(\widetilde{A}) \leq \quad \lambda_1(\widetilde{M}) + \delta_{\max},$$

avec

$$\begin{aligned}
\delta_{\min} &= \frac{1}{2}\left(\|P_{\langle(V_1,V_2)\rangle}(x)\|_2^2 - gap_2 + \sqrt{(gap_2 + \|P_{\langle(V_1,V_2)\rangle}(x)\|_2^2)^2 - 4\ gap_2\|P_{\langle(V_2)\rangle}(x)\|_2^2}\right)\\
\delta_{\max} &= \frac{1}{2}\left(\|x\|_2^2 - gap_2 + \sqrt{(gap_2 + \|x\|_2^2)^2 - 4\ gap_2\|P_{\langle(V_2,\ldots,V_d)\rangle}(x)\|_2^2}\right),
\end{aligned}$$

où $(V_i, \ldots, V_j)$, $1 \leq i \leq j \leq d$, denote tl'espace vectoriel engendré par $V_i, \ldots, V_j$ et $P_{\langle(V_i,\ldots,V_j)\rangle}$ denote la projection othogonale sur cet espace, et

$$gap_2 \quad = \quad \lambda_1(\widetilde{M}) - \lambda_2(\widetilde{M}).$$

Le problème de cette inégalité est qu'ell fait intervenir le trou spectral, qui n'est pas toujours facile à estimer dans les applications.

Resultats obtenus

Notre résultat principal est le théorème suivant, où un encadrement relativement simple est proposé pour la valeur propre la plus grande de la matrice perturbée.

Theorem 6.5 *Soit $M \in \mathbb{C}^{d\times d}$ une matrice hermitienne, dont les valeur propres sont notées $\lambda_1 \geq \cdots \geq \lambda_d$ et leurs vecteurs propres correspondants $(V_1, \cdots, V_d)$. Soient $c \in \mathbb{R}$, $a \in \mathbb{C}^d$. Soit A donné par (1.1). Alors:*

$$\frac{2\langle a, V_1\rangle^2}{\eta_1 + \sqrt{\eta_1^2 + 4\langle a, V_1\rangle^2}} \leq \lambda_1(A) - \max(c, \lambda_1) \leq \frac{2\|a\|^2}{\eta_1 + \sqrt{\eta_1^2 + 4\|a\|^2}}, \tag{6.78}$$

avec

$$\eta_1 \quad = \quad |c - \lambda_1|.$$

Remark 6.6

- *L'inégalité (3.1) est assez précise: en effet, elle est atteinte pour $M = I$, $c = 1$ et pour tout a, tel que $\lambda_{\max}(A) = 1 + \|a\|$;*
- *La borne (3.1) est meilleure que (2.-2) car $\eta_1 \geq \eta$. Un exemple typique où on voit bien l'amélioration est quand c est l'une des valeurs propres de M (i.e. $\eta = 0$). Par exemple, prenons $c = 1$, $a^t = (\alpha, 0)$ et $M = diag(2, 1)$. En particulier, $\eta_1 = 1$. Un calcul rapide montre que $\lambda_1(A) = 3/2 + \sqrt{1/4 + \alpha^2}$ et donc*

$$\lambda_1(A) - \lambda_1(M) = \sqrt{1/4 + \alpha^2} - 1/2 = \frac{2\alpha^2}{1 + \sqrt{1 + 4\alpha^2}},$$

  qui est la borne supérieure de (3.1), alors que la borne (2.-2) est simplement donnée par l'inégalité triangulaire $|\lambda_1(A) - \lambda_1(M)| \leq |\alpha|$.

- *La borne inférieure dans (3.1) est aussi meilleure que celle de (2.-2) car:*

$$\lambda_1(A) \geq \max(c, \lambda_1) + \frac{2\langle a, V_1\rangle^2}{\eta_1 + \sqrt{\eta_1^2 + 4\langle a, V_1\rangle^2}} \geq \max(c, \lambda_1) - \frac{2\|a\|^2}{\eta + \sqrt{\eta^2 + 4\|a\|^2}}.$$

  En particulier, notre borne inférieure est compatible avec le théorème d'entrelacement de Cauchy, qui dit que $\lambda_1(A) \geq \lambda_1$.

Polynomes de Laguerre et matrices de covariance

Nous nous sommes penchés alors sur l'étude des matrices de covariances. En particulier, nous avons été vite aspirés par l'idée que si l'on connaît les espacements entres les valeurs propres d'une matrice de covariance, on n'en pourra que mieux contrôler les perturbations obtenues par l'apparition d'une nouvelle donnée. C'est ce type d'approche incrémentale toute bête qui est mise en oeuvre de manière très technique et subtile par Srivastava et Vershynin dans leur étude récente des matrices de covariance, étude surpassée depuis par les résultats sus-cité de Koltchinski et Mendelson. Cela s'est averé une expérience très enrichissante. Nous avons réalisé que peu de choses sont connues sur les espacements entre valeurs propres successives pour beaucoup de modèles naturels en probabilité. Un travail impressionnant de P. Bourgade et G. Ben Arous a donné des résultats asymptotiques sur l' espacement le plus grand et le plus petit. De notre côté, nous étions intéressés par des résultats non-asymptotiques dans l'esprit des travaux de Rudelson et Vershynin sur les valeurs propres extrêmes. Un résultat nous a vite interpelé: le mode de la loi de Whishart est donné par le

vecteur des zéros d'un polynôme de Laguerre. Ce résultat trouvé dans un polycopié de J. Faraut sur les matrices aléatoires et retrouvé dans les diverses publication de Holger Dette sur les polynômes orthogonaux, nous a incité à regarder les espacements successifs entre les zéros des polynômes de Laguerre. Le résultat que nous avons obtenu, par une méthode simple qui s'appuie sur le Bethe Ansatz que nous avions découvert dans un article de Krasikov, est une borne inférieure uniforme sur ces espacement. Elle est suffisamment simple, et de démonstration élémentaire, pour avoir été jugée favorablement, à notre grande surprise, pour publication dans les proceedings de l'AMS, journal que nous avons choisi par curiosité, après avoir été rejeté par Applied Mathematics Letters, appuyé du commentaire laconique "too theoretical". Nous nous attendions à ce que l'article soit rejeté pour des raisons enfin plus argumentées, sur un sujet vieux comme Gauss, et sur lequel nous ne sommes encore que de naifs spectateurs gourmands de savoir. Par chance pour notre H-index, ce que nous avons fait n'avait pas été regardé auparavent, et avec autant de naiveté, par les experts du domaine. L'article fait l'objet de la Section M.

Rappelons un résultat très intéressant, que l'on peut trouver dans le Lemme 1 de [11]. Soit f un polynome à coefficients réels avec des zéros simples $x_1 < \cdots < x_n$, satisfaisant l'ODE $f'' - 2af' + bf = 0$ où a et b sont des fonctions mŕomorphes dont les pôles sont différents des x_i's. Alors, pour tout $k \in \{1 \cdots n\}$ fixé,

$$\sum_{j \neq k} \frac{1}{(x_k - x_j)^2} \quad = \quad \frac{\Delta(x_k) - 2a'(x_k)}{3}, \tag{6.79}$$

avec $\Delta(x) = b(x) - a^2(x)$. Ce type d'égalités est appelé équations de "Bethe *ansatz*" en anglais.

Pour $\alpha > -1$, les polynômes de Laguerre $L_n^{(\alpha)}$ (n indique le degré) sont des polynômes orthogonaux par rapport aux poids $x^\alpha e^{-x}$ sur $(0, \infty)$. Soient $x_{n,n}(\alpha) < \cdots < x_{n,1}(\alpha)$ les zeros de $L_n^{(\alpha)}$. On sait depuis longtemps que $L_n^{(\alpha)}$ est une solution de l'équation différentielle:

$$u'' - \left(1 - \frac{\alpha+1}{x}\right) u' + \frac{n}{x} u \quad = \quad 0.$$

Cela donne $a(x) = \frac{1}{2}\left(1 - \frac{\alpha+1}{x}\right)$. Ainsi,

$$\Delta(x) = \frac{n}{x} - \frac{(x-\alpha-1)^2}{4x^2} = \frac{-x^2 + (2(\alpha+1)+4n)x - (\alpha+1)^2}{4x^2},$$

et donc,

$$\Delta(x) \quad = \quad \frac{(U^2 - x)(x - V^2)}{4x^2}, \tag{6.80}$$

où

$$U = \sqrt{n+\alpha+1} + \sqrt{n}, \quad V = \sqrt{n+\alpha+1} - \sqrt{n}. \tag{6.81}$$

Comme le membre de gauche de (2.1) est positif et $a'(x) > 0$ pour $x > 0$, une conséquence immédiate de (2.1) est que pour tout k, $(U^2 - x_{n,k}(\alpha))(x_{n,k}(\alpha) - V^2) > 0$, i.e.

$$V^2 < x_{n,n}(\alpha) < x_{n,1}(\alpha) < U^2. \tag{6.82}$$

De nombreuses bornes sur les zeros extrêmes sont connues. On peut consulter par exemple [4, 7, 10, 11, 13]. Utilisant également le Bethe *ansatz*, Krasikov a prouvé dans [11, Theorem 1]:

$$V^2 + 3V^{4/3}(U^2 - V^2)^{-1/3} \leq x_{n,n}(\alpha) < x_{n,1}(\alpha) \leq U^2 - 3U^{4/3}(U^2 - V^2)^{-1/3} + 2. \tag{6.83}$$

Notre résultat principal est le théorème suivant.

Theorem 6.7 *On suppose $\alpha > -1$. Alors, pour tout $k \in \{1, \cdots, n-1\}$, les espacements successifs entre les zéros admettent la borne inférieure uniforme suivante:*

$$x_{n,k}(\alpha) - x_{n,k+1}(\alpha) \quad \geq \quad \sqrt{3}\frac{\alpha+1}{\sqrt{n(n+\alpha+1)}}. \tag{6.84}$$

De puis, si $\alpha \geq n/C$ pour un certain $C > 0$, on a

$$x_{n,k}(\alpha) - x_{n,k+1}(\alpha) \quad \geq \quad \frac{1}{\sqrt{C+1}}\sqrt{\frac{\alpha}{n}}. \tag{6.85}$$

Chapter A

Kullback Proximal Algorithms for Maximum Likelihood Estimation

with Alfred O. Hero.

Abstract

Accelerated algorithms for maximum likelihood image reconstruction are essential for emerging applications such as 3D tomography, dynamic tomographic imaging, and other high dimensional inverse problems. In this paper, we introduce and analyze a class of fast and stable sequential optimization methods for computing maximum likelihood estimates and study its convergence properties. These methods are based on a proximal point algorithm implemented with the Kullback-Liebler (KL) divergence between posterior densities of the complete data as a proximal penalty function. When the proximal relaxation parameter is set to unity one obtains the classical expectation maximization (EM) algorithm. For a decreasing sequence of relaxation parameters, relaxed versions of EM are obtained which can have much faster asymptotic convergence without sacrifice of monotonicity. We present an implementation of the algorithm using Moré's Trust Region update strategy. For illustration the method is applied to a non-quadratic inverse problem with Poisson distributed data.

1 Introduction

Maximum likelihood (ML) or maximum penalized likelihood (MPL) approaches have been widely adopted for image restoration and image reconstruction from noise contaminated data with known statistical distribution. In many cases the likelihood function is in a form for which analytical solution is difficult or impossible. When this is the case iterative solutions to the ML reconstruction or restoration problem are of interest. Among the most stable iterative strategies for ML is the popular expectation maximization (EM) algorithm [8]. The EM algorithm has been widely applied to emission and transmission computed tomography [39, 23, 36] with Poisson data. The EM algorithm has the attractive property of monotonicity which guarantees that the likelihood function increases with each iteration. The convergence properties of the EM algorithm and its variants have been extensively studied in the literature; see [42] and [15] for instance. It is well known that under strong concavity assumptions the EM algorithm converges linearly towards the ML estimator θ_{ML}. However, the rate coefficient is small and in practice the EM algorithm suffers from slow

convergence in late iterations. Efforts to improve on the asymptotic convergence rate of the EM algorithm have included: Aitken's acceleration [28], over-relaxation [26], conjugate gradient [20] [19], Newton methods [30] [4], quasi-Newton methods [22], ordered subsets EM [17] and stochastic EM [25]. Unfortunately, these methods do not automatically guarantee the monotone increasing likelihood property as does standard EM. Furthermore, many of these accelerated algorithms require additional monitoring for instability [24]. This is especially problematic for high dimensional image reconstruction problems, e.g. 3D or dynamic imaging, where monitoring could add significant computational overhead to the reconstruction algorithm.

The contribution of this paper is the introduction of a class of accelerated EM algorithms for likelihood function maximization via exploitation of a general relation between EM and proximal point (PP) algorithms. These algorithms converge and can have quadratic rates of convergence even with approximate updating. Proximal point algorithms were introduced by Martinet [5] and Rockafellar [9], based on the work of Minty [31] and Moreau [33], for the purpose of solving convex minimization problems with convex constraints. A key motivation for the PP algorithm is that by adding a sequence of iteration-dependent penalties, called proximal penalties, to the objective function to be maximized one obtains stable iterative algorithms which frequently outperform standard optimization methods without proximal penalties, e.g. see Goldstein and Russak [1]. Furthermore, the PP algorithm plays a paramount role in non-differentiable optimization due to its connections with the Moreau-Yosida regularization; see Minty [31], Moreau [33], Rockafellar [9] and Hiriart-Hurruty and Lemaréchal [16].

While the original PP algorithm used a simple quadratic penalty more general versions of PP have recently been proposed which use non-quadratic penalties, and in particular entropic penalties. Such penalties are most commonly applied to ensure non-negativity when solving Lagrange duals of inequality constrained primal problems; see for example papers by Censor and Zenios [5], Ekstein [10], Eggermont [9], and Teboulle [19]. In this paper we show that by choosing the proximal penalty function of PP as the Kullback-Liebler (KL) divergence between successive iterates of the posterior densities of the complete data, a generalization of the generic EM maximum likelihood algorithm is obtained with accelerated convergence rate. When the relaxation sequence is constant and equal to unity the PP algorithm with KL proximal penalty reduces to the standard EM algorithm. On the other hand for a decreasing relaxation sequence the PP algorithm with KL proximal penalty is shown to yield an iterative ML algorithm which has much faster convergence than EM without sacrificing its monotonic likelihood property.

It is important to point out that relations between particular EM and particular PP algorithms have been previously observed, but not in the full generality established in this paper. Specifically, for parameters constrained to the non-negative orthant, Eggermont [9] established a relation between an entropic modification of the standard PP algorithm and a class of multiplicative methods for smooth convex optimization. The modified PP algorithm that was introduced in [9] was obtained by replacing the standard quadratic penalty by the relative entropy between successive non-negative parameter iterates. This extension was shown to be equivalent to an "implicit" algorithm which, after some approximations to the exact PP objective function, reduces to the "explicit" Shepp and Vardi EM algorithm [39] for image reconstruction in emission tomography. Eggermont [9] went on to prove that the explicit and implicit algorithms are monotonic and both converge when the sequence of relaxation parameters is bounded below by a strictly positive number.

In contrast to [9], here we establish a general and exact relation between the generic EM procedure, i.e. arbitrary incomplete and complete data distributions, and an extended class of PP algorithms. As pointed out above, the extended PP algorithm is implemented

with a proximal penalty which is the relative entropy (KL divergence) between successive iterates of the posterior densities of the complete data. This modification produces a class of algorithms which we refer to as Kullback-Liebler proximal point (KPP). We prove a global convergence result for the KPP algorithm under strict concavity assumptions. An approximate KPP is also proposed using the Trust Region strategy [32, 34] adapted to KPP. We show, in particular, that both the exact and approximate KPP algorithms have superlinear convergence rates when the sequence of positive relaxation parameters converge to zero. Finally, we illustrate these results for KPP acceleration of the Shepp and Vardi EM algorithm implemented with Trust Region updating.

The results given here are also applicable to the non-linear updating methods of Kivinen and Warmuth [21] for accelerating the convergence of Gaussian mixture-model identification algorithms in supervised machine learning, see also Warmuth and Azoury [41] and Helmbold, Schapire, Singer and Warmuth [14]. Indeed, similarly to the general KPP algorithm introduced in this paper, in [14] the KL divergence between the new and the old mixture model was added to the gradient of the Gaussian mixture-model likelihood function, appropriately weighted with a multiplicative factor called the learning rate parameter. This procedure led to what the authors of [14] called an exponentiated gradient algorithm. These authors provided experimental evidence of significant improvements in convergence rate as compared to gradient descent and ordinary EM. The results in this paper provide a general theory which validate such experimental results for a very broad class of parametric estimation problems.

The outline of the paper is as follows. In Section 2 we provide a brief review of key elements of the classical EM algorithm. In Section 2, we establish the general relationship between the EM algorithm and the proximal point algorithm. In section 4, we present the general KPP algorithm and we establish global and superlinear convergence to the maximum likelihood estimator for a smooth and strictly concave likelihood function. In section 5, we study second order approximations of the KPP iteration using Trust Region updating. Finally, in Section 6 we present numerical comparisons for a Poisson inverse problem.

2 Background

The problem of maximum likelihood (ML) estimation consists of finding a solution of the form

$$\theta_{ML} = \text{argmax}_{\theta \in \mathbb{R}^p} \; l_y(\theta), \tag{2.1}$$

where y is an observed sample of a random variable Y defined on a sample space $\mathcal{Y}$ and $l_y(\theta)$ is the log-likelihood function defined by

$$l_y(\theta) = \log g(y; \theta), \tag{2.2}$$

and $g(y; \theta)$ denotes the density of Y at y parametrized by a vector parameter θ in $\mathbb{R}^p$. One of the most popular iterative methods for solving ML estimation problems is the Expectation Maximization (EM) algorithm described in Dempster, Laird, and Rubin [8] which we recall for the reader.

A more informative data space $\mathcal{X}$ is introduced. A random variable X is defined on $\mathcal{X}$ with density $f(x; \theta)$ parametrized by θ. The data X is more informative than the actual data Y in the sense that Y is a compression of X, i.e. there exists a non-invertible transformation h such that $Y = h(X)$. If one had access to the data X it would therefore be advantageous to replace the ML estimation problem (3.1) by

$$\hat{\theta}_{ML} = \text{argmax}_{\theta \in \mathbb{R}^p} l_x(\theta), \tag{2.3}$$

with $l_x(\theta) = \log f(x;\theta)$. Since $y = h(x)$ the density g of Y is related to the density f of X through

$$g(y;\theta) = \int_{h^{-1}(\{y\})} f(x;\theta) d\mu(x) \tag{2.4}$$

for an appropriate measure μ on $\mathcal{X}$. In this setting, the data y are called *incomplete data* whereas the data x are called *complete data*.

Of course the complete data x corresponding to a given observed sample y are unknown. Therefore, the complete data likelihood function $l_x(\theta)$ can only be estimated. Given the observed data y and a previous estimate of θ denoted $\bar{\theta}$, the following minimum mean square error estimator (MMSE) of the quantity $l_x(\theta)$ is natural

$$Q(\theta, \bar{\theta}) = \mathbb{E}[\log f(x;\theta)|y;\bar{\theta}],$$

where, for any integrable function $F(x)$ on $\mathcal{X}$, we have defined the conditional expectation

$$\mathbb{E}[F(x)|y;\bar{\theta}] = \int_{h^{-1}(\{y\})} F(x) k(x|y;\bar{\theta}) d\mu(x)$$

and $k(x|y;\bar{\theta})$ is the conditional density function given y

$$k(x|y;\bar{\theta}) = \frac{f(x;\bar{\theta})}{g(y;\bar{\theta})}. \tag{2.5}$$

The EM algorithm generates a sequence of approximations to the solution (2.4) starting from an initial guess θ^0 of θ_{ML} and is defined by

$$\begin{aligned} &\text{Compute } Q(\theta, \theta^k) = \mathbb{E}[\log f(x;\theta)|y;\theta^k] && \text{E Step} \\ &\theta^{k+1} = \operatorname{argmax}_{\theta\in\mathbb{R}^p} Q(\theta, \theta^k) && \text{M Step} \end{aligned}$$

A key to understanding the convergence of the EM algorithm is the decomposition of the likelihood function presented in Dempster, Laird and Rubin [8]. As this decomposition is also the prime motivation for the KPP generalization of EM it will be worthwhile to recall certain elements of their argument. The likelihood can be decomposed as

$$l_y(\theta) = Q(\theta, \bar{\theta}) + H(\theta, \bar{\theta}) \tag{2.6}$$

where

$$H(\theta, \bar{\theta}) = -\mathbb{E}[\log k(x|y;\theta)|y;\bar{\theta}].$$

It follows from elementary application of Jensen's inequality to the log function that

$$H(\theta, \bar{\theta}) \geqslant H(\theta, \theta) \geqslant 0, \quad \forall \theta,\ \bar{\theta} \in \mathbb{R}^p. \tag{2.7}$$

Observe from (2.6) and (2.7) that for any θ^k the θ function $Q(\theta, \theta^k)$ is a lower bound on the log likelihood function $l_y(\theta)$. This property is sufficient to ensure monotonicity of the algorithm. Specifically, since the the M-step implies that

$$Q(\theta^{k+1}, \theta^k) \geqslant Q(\theta^k, \theta^k), \tag{2.8}$$

one obtains

$$\begin{aligned} l_y(\theta^{k+1}) - l_y(\theta^k) \quad \geqslant \quad & Q(\theta^{k+1}, \theta^k) - Q(\theta^k, \theta^k) \\ & + H(\theta^{k+1}, \theta^k) - H(\theta^k, \theta^k). \end{aligned} \tag{2.9}$$

Hence, using (2.8) and (2.7)

$$l_y(\theta^{k+1}) \geqslant l_y(\theta^k).$$

This is the well known monotonicity property of the EM algorithm.

Note that if the function $H(\theta, \bar{\theta})$ in (2.6) were scaled by an arbitrary positive factor β the function $Q(\theta, \bar{\theta})$ would remain a lower bound on $l_y(\theta)$, the right hand side of (2.9) would remain positive and monotonicity of the algorithm would be preserved. As will be shown below, if β is allowed to vary with iteration in a suitable manner one obtains a monotone, superlinearly convergent generalization of the EM algorithm.

3 Proximal point methods and the EM algorithm

In this section, we present the proximal point (PP) algorithm of Rockafellar and Martinet. We then demonstrate that EM is a particular case of proximal point implemented with a Kullback-type proximal penalty.

The proximal point algorithm

Consider the general problem of maximizing a concave function $\Phi(\theta)$. The proximal point algorithm is an iterative procedure which can be written

$$\theta^{k+1} = \text{argmax}_{\theta \in \mathbb{R}^p} \left\{ \Phi(\theta) - \frac{\beta_k}{2} \|\theta - \theta^k\|^2 \right\}. \tag{3.10}$$

The quadratic penalty $\|\theta - \theta^k\|^2$ is relaxed using a sequence of positive parameters $\{\beta_k\}$. In [9], Rockafellar showed that superlinear convergence of this method is obtained when the sequence $\{\beta_k\}$ converges towards zero. In numerical implementations of proximal point the function $\Phi(\theta)$ is generally replaced by a piecewise linear model [16].

Proximal interpretation of the EM algorithm

In this section, we establish an exact relationship between the generic EM procedure and an extended proximal point algorithm. For our purposes, we will need to consider a particular Kullback-Liebler (KL) information measure. Assume that the family of conditional densities $\{k(x|y;\theta)\}_{\theta \in \mathbb{R}^p}$ is regular in the sense of Ibragimov and Khasminskii [9], in particular $k(x|y;\theta)\mu(x)$ and $k(x|y;\bar{\theta})\mu(x)$ are mutually absolutely continuous for any θ and $\bar{\theta}$ in $\mathbb{R}^p$. Then the Radon-Nikodym derivative $\frac{k(x|y,\bar{\theta})}{k(x|y;\theta)}$ exists for all $\theta, \bar{\theta}$ and we can define the following KL divergence:

$$I_y(\bar{\theta}, \theta) = \mathbb{E}\big[\log \frac{k(x|y,\bar{\theta})}{k(x|y;\theta)} | y; \bar{\theta}\,\big]. \tag{3.11}$$

Proposition 3.1 *The EM algorithm is equivalent to the following recursion with* $\beta_k = 1$, $k = 1, 2, \ldots,$

$$\theta^{k+1} = argmax_{\theta \in \mathbb{R}^p} \left\{ l_y(\theta) - \beta_k I_y(\theta^k, \theta) \right\} \tag{3.12}$$

For general positive sequence $\{\beta_k\}$ the recursion in Proposition 2.1 can be identified as a modification of the PP algorithm (2.5) with the standard quadratic penalty replaced by the KL penalty (2.4) and having relaxation sequence $\{\beta_k\}$. In the sequel we call this modified

PP algorithm the Kullback-Liebler proximal point (KPP) algorithm. In many treatments of the EM algorithm the quantity

$$Q(\theta, \bar{\theta}) = l_y(\theta) - l_y(\bar{\theta}) - I(\bar{\theta}, \theta)$$

is the surrogate function that is maximized in the M-step. This surrogate objective function is identical (up to an additive constant) to the KPP objective $l_y(\theta) - \beta_k I_y(\theta^k, \theta)$ of (3.12) when $\beta_k = 1$.

Proof of Proposition 2.1: The key to making the connection with the proximal point algorithm is the following representation of the M step:

$$\theta^{k+1} = \text{argmax}_{\theta \in \mathbb{R}^p} \big\{ \log g(y;\theta) + \mathbb{E}\big[\log \frac{f(x;\theta)}{g(y;\theta)} | y; \theta^k\big] \big\}.$$

This equation is equivalent to

$$\theta^{k+1} = \text{argmax}_{\theta \in \mathbb{R}^p} \big\{ \log g(y;\theta) + \mathbb{E}\big[\log \frac{f(x;\theta)}{g(y;\theta)} | y; \theta^k\big] - \mathbb{E}\big[\log \frac{f(x;\theta^k)}{g(y;\theta^k)} | y; \theta^k\big] \big\}$$

since the additional term is constant in θ. Recalling that $k(x|y;\theta) = \frac{f(x;\theta)}{g(y;\theta)}$,

$$\theta^{k+1} = \text{argmax}_{\theta \in \mathbb{R}^p} \big\{ \log g(y;\theta) + \mathbb{E}\big[\log k(x|y;\theta) | y; \theta^k\big] - \mathbb{E}\big[\log k(x|y;\theta^k) | y; \theta^k\big] \big\}.$$

We finally obtain

$$\theta^{k+1} = \text{argmax}_{\theta \in \mathbb{R}^p} \big\{ \log g(y;\theta) + \mathbb{E}\big[\log \frac{k(x|y;\theta)}{k(x|y;\theta^k)} | y; \theta^k\big] \big\}$$

which concludes the proof.

4 Convergence of the KPP Algorithm

In this section we establish monotonicity and other convergence properties of the KPP algorithm of Proposition 2.1.

Monotonicity

For bounded domain of θ, the KPP algorithm is well defined since the maximum in (3.12) is always achieved in a bounded set. Monotonicity is guaranteed by this procedure as proved in the following proposition.

Proposition 4.1 *The log-likelihood sequence $\{l_y(\theta^k)\}$ is monotone non-decreasing and satisfies*

$$l_y(\theta^{k+1}) - l_y(\theta^k) \geqslant \beta_k I_y(\theta^k, \theta^{k+1}), \tag{4.13}$$

Proof: From the recurrence in (3.12), we have

$$l_y(\theta^{k+1}) - l_y(\theta^k) \geqslant \beta_k I_y(\theta^k, \theta^{k+1}) - \beta_k I_y(\theta^k, \theta^k).$$

Since $I_y(\theta^k, \theta^k) = 0$ and $I_y(\theta^k, \theta^{k+1}) \geqslant 0$, we deduce (4.13) and that $\{l_y(\theta^k)\}$ is non-decreasing.

We next turn to asymptotic convergence of the KPP iterates $\{\theta^k\}$.

Asymptotic Convergence

In the sequel $\nabla_{01} I_y(\bar{\theta}, \theta)$ (respectively $\nabla^2_{01} I_y(\bar{\theta}, \theta)$) denotes the gradient (respectively the Hessian matrix) of $I_y(\bar{\theta}, \theta)$ in the first variable. For a square matrix M, Λ_M denotes the greatest eigenvalue of a matrix M and λ_M denotes the smallest.

We make the following assumptions

Assumptions 4.1 *We assume the following:*

(i) $l_y(\theta)$ *is twice continuously differentiable on* $\mathbb{R}^p$ *and* $I_y(\bar{\theta}, \theta)$ *is twice continuously differentiable in* $(\theta, \bar{\theta})$ *in* $\mathbb{R}^p \times \mathbb{R}^p$.

(ii) $\lim_{\|\theta\| \to \infty} l_y(\theta) = -\infty$ *where* $\|\theta\|$ *is the standard Euclidean norm on* $\mathbb{R}^p$.

(iii) $l_y(\theta) < \infty$ *and* $\Lambda_{\nabla^2 l_y(\theta)} < 0$ *on every bounded* θ*-set.*

(iv) for any $\bar{\theta}$ *in* $\mathbb{R}^p$, $I_y(\bar{\theta}, \theta) < \infty$ *and* $0 < \lambda_{\nabla^2_{01} I_y(\bar{\theta}, \theta)} \leqslant \Lambda_{\nabla^2_{01} I_y(\bar{\theta}, \theta)}$ *on every bounded* θ*-set.*

These assumptions ensure smoothness of $l_y(\theta)$ and $I_y(\bar{\theta}, \theta)$ and their first two derivatives in θ. Assumption 5.1.iii also implies strong concavity of $l_y(\theta)$. Assumption 5.1.iv implies that $I_y(\bar{\theta}, \theta)$ is strictly convex and that the parameter θ is strongly identifiable in the family of densities $k(x|y;\theta)$ (see proof of Lemma 4.3 below). Note that the above assumptions are not the minimum possible set, e.g. that $l_y(\theta)$ and $I_y(\bar{\theta}, \theta)$ are upper bounded follows from continuity, Assumption 5.1.ii and the property $I_y(\bar{\theta}, \theta) \geqslant I_y(\bar{\theta}, \bar{\theta}) = 0$, respectively.

We first characterize the fixed points of the KPP algorithm.

A result that will be used repeatedly in the sequel is that for any $\bar{\theta} \in \mathbb{R}^p$

$$\nabla_{01} I_y(\bar{\theta}, \bar{\theta}) = 0. \tag{4.14}$$

This follows immediately from the information inequality for the KL divergence [6, Thm. 2.6.3]

$$I_y(\bar{\theta}, \theta) \geqslant I_y(\bar{\theta}, \bar{\theta}) = 0,$$

so that, by smoothness Assumption 5.1.i, $I_y(\bar{\theta}, \theta)$ has a stationary point at $\theta = \bar{\theta}$.

Proposition 4.2 *Let the densities* $g(y;\theta)$ *and* $k(x|y;\theta)$ *be such that Assumptions 5.1 are satisfied. Then the fixed points of the recurrence in (3.12) are maximizers of the log-likelihood function* $l_y(\theta)$ *for any relaxation sequence* $\beta_k = \beta > 0$, $k = 1, 2, \ldots$.

Proof: Consider a fixed point θ^* of the recurrence relation (3.12) for $\beta_k = \beta =$ constant. Then,

$$\theta^* = \operatorname{argmax}_{\theta \in \mathbb{R}^p} \left\{ l_y(\theta) - \beta I_y(\theta^*, \theta) \right\}.$$

As $l_y(\theta)$ and $I_y(\theta^*, \theta)$ are both smooth in θ, θ^* must be a stationary point

$$0 = \nabla l_y(\theta^*) - \beta \nabla_{01} I_y(\theta^*, \theta^*).$$

Thus, as by (4.14) $\nabla_{01} I_y(\theta^*, \theta^*) = 0$,

$$0 = \nabla l_y(\theta). \tag{4.15}$$

Since $l_y(\theta)$ is strictly concave, we deduce that θ^* is a maximizer of $l_y(\theta)$.

The following will be useful.

Lemma 4.3 *Let the conditional density $k(x|y;\theta)$ be such that $I_y(\bar{\theta},\theta)$ satisfies Assumption 5.1.iv. Then, given two bounded sequences $\{\theta_1^k\}$ and $\{\theta_2^k\}$, $\lim_{k\to\infty} I_y(\theta_1^k,\theta_2^k)=0$ implies that $\lim_{k\to\infty}\|\theta_1^k-\theta_2^k\|=0$.*

Proof: Let $\mathcal{B}$ be any bounded set containing both sequences $\{\theta_1^k\}$ and $\{\theta_2^k\}$. Let λ denote the minimum

$$\lambda=\min_{\theta,\bar{\theta}\in\mathcal{B}}\lambda_{\nabla_{01}^2 I_y(\bar{\theta},\theta)} \tag{4.16}$$

Assumption 5.1.iv implies that $\lambda>0$. Furthermore, invoking Taylor's theorem with remainder, $I_y(\bar{\theta},\theta)$ is strictly convex in the sense that for any k

$$\begin{aligned} I_y(\theta_1^k,\theta_2^k)\geqslant I_y(\theta_1^k,\theta_1^k)+\quad & \nabla I_y(\theta_1^k,\theta_1^k)^T(\theta_1^k-\theta_2^k)\\ & +\tfrac{1}{2}\lambda\|\theta_1^k-\theta_2^k\|^2. \end{aligned}$$

As $I_y(\theta_1^k,\theta_1^k)=0$ and $\nabla_{01}I_y(\theta_1^k,\theta_1^k)=0$, recall (4.14), we obtain

$$I_y(\theta_1^k,\theta_2^k)\geqslant\frac{\lambda}{2}\|\theta_1^k-\theta_2^k\|^2.$$

The desired result comes from passing to the limit $k\to\infty$.

Using these results, we easily obtain the following.

Lemma 4.4 *Let the densities $g(y;\theta)$ and $k(x|y;\theta)$ be such that Assumptions 5.1 are satisfied. Then $\{\theta^k\}_{k\in\mathbb{N}}$ is bounded.*

Proof: Due to Proposition 5.44, the sequence $\{l_y(\theta^k)\}$ is monotone increasing. Therefore, assumption 5.1.ii implies that $\{\theta^k\}$ is bounded.

In the following lemma, we prove a result which is often called asymptotic regularity [13].

Lemma 4.5 *Let the densities $g(y;\theta)$ and $k(x|y;\theta)$ be such that $l_y(\theta)$ and $I_y(\bar{\theta},\theta)$ satisfy Assumptions 5.1. Let the sequence of relaxation parameters $\{\beta_k\}_{k\in\mathbb{N}}$ satisfy $0<\liminf\beta_k\leqslant\limsup\beta_k<\infty$. Then,*

$$\lim_{k\to\infty}\|\theta^{k+1}-\theta^k\|=0. \tag{4.17}$$

Proof: By Assumption 5.1.iii and by Proposition 5.44 $\{l_y(\theta^k)\}_{k\in\mathbb{N}}$ is bounded and monotone. Since, by Lemma 4.4, $\{\theta^k\}_{k\in\mathbb{N}}$ is a bounded sequence $\{l_y(\theta^k)\}_{k\in\mathbb{N}}$ converges. Therefore, $\lim_{k\to\infty}\{l_y(\theta^{k+1})-l_y(\theta^k)\}=0$ which, from (4.13), implies that $\beta_k I_y(\theta^k,\theta^{k+1})$ vanishes when k tends to infinity. Since $\{\beta_k\}_{k\in\mathbb{N}}$ is bounded below by $\liminf\beta_k>0$: $\lim_{k\to\infty}I_y(\theta^k,\theta^{k+1})=0$. Therefore, Lemma 4.3 establishes the desired result.

We can now give a global convergence theorem.

Theorem 4.6 *Let the sequence of relaxation parameters $\{\beta_k\}_{k\in\mathbb{N}}$ be positive and converge to a limit $\beta^*\in[0,\infty)$. Then the sequence $\{\theta^k\}_{k\in\mathbb{N}}$ converges to the solution of the ML estimation problem (3.1).*

Proof: Since $\{\theta^k\}_{k\in\mathbb{N}}$ is bounded, one can extract a convergent subsequence $\{\theta^{\sigma(k)}\}_{k\in\mathbb{N}}$ with limit θ^*. The defining recurrence (3.12) implies that

$$\nabla l_y(\theta^{\sigma(k)+1})-\beta_{\sigma(k)}\nabla_{01}I_y(\theta^{\sigma(k)},\theta^{\sigma(k)+1})=0. \tag{4.18}$$

We now prove that θ^* is a stationary point of $l_y(\theta)$. Assume first that $\{\beta_k\}_{k\in\mathbb{N}}$ converges to zero, i.e. $\beta^* = 0$. Due to Assumptions 5.1.i, $\nabla l_y(\theta)$ is continuous in θ. Hence, since $\nabla_{01} I_y(\bar{\theta}, \theta)$ is bounded on bounded subsets, (4.18) implies

$$\nabla l_y(\theta^*) = 0.$$

Next, assume that $\beta^* > 0$. In this case, Lemma 4.5 establishes that

$$\lim_{k\to\infty} \|\theta^{k+1} - \theta^k\| = 0.$$

Therefore, $\{\theta^{\sigma(k)+1}\}_{k\in\mathbb{N}}$ also tends to θ^*. Since $\nabla_{01} I_y(\bar{\theta}, \theta)$ is continuous in $(\bar{\theta}, \theta)$ equation (4.18) gives at infinity

$$\nabla l_y(\theta^*) - \beta^* \nabla_{01} I_y(\theta^*, \theta^*) = 0.$$

Finally, by (4.14), $\nabla_{01} I_y(\theta^*, \theta^*) = 0$ and

$$\nabla l_y(\theta^*) = 0. \tag{4.19}$$

The proof is concluded as follows. As, by Assumption 5.1.iii, $l_y(\theta)$ is concave, θ^* is a maximizer of $l_y(\theta)$ so that θ^* solves the Maximum Likelihood estimation problem (3.1). Furthermore, as positive definiteness of $\nabla^2 l_y$ implies that $l_y(\theta)$ is in fact strictly concave, this maximizer is unique. Hence, $\{\theta^k\}$ has only one accumulation point and $\{\theta^k\}$ converges to θ^* which ends the proof.

We now establish the main result concerning speed of convergence. Recall that a sequence $\{\theta^k\}$ is said to converge superlinearly to a limit θ^* if:

$$\lim_{k\to\infty} \frac{\|\theta^{k+1} - \theta^*\|}{\|\theta^k - \theta^*\|} = 0, . \tag{4.20}$$

Theorem 4.7 *Assume that the sequence of positive relaxation parameters $\{\beta_k\}_{k\in\mathbb{N}}$ converges to zero. Then, the sequence $\{\theta^k\}_{k\in\mathbb{N}}$ converges superlinearly to the solution of the ML estimation problem (3.1).*

Proof: Due to Theorem 4.6, the sequence $\{\theta^k\}$ converges to the unique maximizer θ_{ML} of $l_y(\theta)$. Assumption 5.1.i implies that the gradient mapping $\nabla_\theta\big(l_y(\theta) - \beta_k I_y(\theta_{ML}, \theta)\big)$ is continuously differentiable. Hence, we have the following Taylor expansion about θ_{ML}.

$$\begin{aligned}\nabla l_y(\theta) - \beta_k \nabla_{01} I_y(\theta_{ML}, \theta) = \nabla l_y(\theta_{ML}) \\ - \beta_k \nabla_{01} I_y(\theta_{ML}, \theta_{ML}) \\ + \nabla^2 l_y(\theta_{ML})(\theta - \theta_{ML}) \\ - \beta_k \nabla_{01}^2 I_y(\theta_{ML}, \theta_{ML})(\theta - \theta_{ML}) \\ + R(\theta - \theta_{ML}),\end{aligned} \tag{4.21}$$

where the remainder satisfies

$$\lim_{\theta\to\theta_{ML}} \frac{\|R(\theta - \theta_{ML})\|}{\|\theta - \theta_{ML}\|} = 0.$$

Since θ_{ML} maximizes $l_y(\theta)$, $\nabla l_y(\theta_{ML}) = 0$. Furthermore, by (4.14), $\nabla_{01} I_y(\theta_{ML}, \theta_{ML}) = 0$. Hence, (4.21) can be simplified to

$$\begin{aligned}\nabla l_y(\theta) - \beta_k \nabla_{01} I_y(\theta_{ML}, \theta) = \nabla^2 l_y(\theta_{ML})(\theta - \theta_{ML}) \\ - \beta_k \nabla_{01}^2 I_y(\theta_{ML}, \theta_{ML})(\theta - \theta_{ML}) + R(\theta - \theta_{ML}).\end{aligned} \tag{4.22}$$

From the defining relation (3.12) the iterate θ^{k+1} satisfies

$$\nabla l_y(\theta^{k+1}) - \beta_k \nabla_{01} I_y(\theta^k, \theta^{k+1}) = 0. \tag{4.23}$$

So, taking $\theta = \theta^{k+1}$ in (4.22) and using (4.23), we obtain

$$\begin{aligned}
&\beta_k\big(\nabla_{01} I_y(\theta^k, \theta^{k+1}) - \nabla_{01} I_y(\theta_{ML}, \theta^{k+1})\big) = \\
&+ \nabla^2 l_y(\theta_{ML})(\theta^{k+1} - \theta_{ML}) - \beta_k \nabla^2_{01} I_y(\theta_{ML}, \theta_{ML})(\theta^{k+1} - \theta_{ML}) \\
&+ R(\theta^{k+1} - \theta_{ML}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\|\beta_k\big(\nabla_{01} I_y(\theta^k, \theta^{k+1}) - \nabla_{01} I_y(\theta_{ML}, \theta^{k+1})\big) - R(\theta^{k+1} - \theta_{ML})\| = \\
&\|\nabla^2 l_y(\theta_{ML})(\theta^{k+1} - \theta_{ML}) - \beta_k \nabla^2_{01} I_y(\theta_{ML}, \theta_{ML})(\theta^{k+1} - \theta_{ML})\|.
\end{aligned} \tag{4.24}$$

On the other hand, one deduces from Assumptions 5.1 (i) that $\nabla_{01} I_y(\bar\theta, \theta)$ is locally Lipschitz in the variables θ and $\bar\theta$. Then, since, $\{\theta^k\}$ is bounded, there exists a bounded set $\mathcal{B}$ containing $\{\theta^k\}$ and a finite constant L such that for all θ, θ', $\bar\theta$ and $\bar\theta'$ in $\mathcal{B}$,

$$\|\nabla_{01} I_y(\bar\theta, \theta) - \nabla_{01} I_y(\bar\theta', \theta')\| \leqslant L\big(\|\theta - \theta'\|^2 + \|\bar\theta - \bar\theta'\|^2\big)^{\frac{1}{2}}.$$

Using the triangle inequality and this last result, (4.24) asserts that for any $\theta \in \mathcal{B}$

$$\begin{aligned}
\beta_k L\|\theta^k - \theta_{ML}\| + \|R(\theta^{k+1} - \theta_{ML})\| \geqslant \|\big(&\nabla^2 l_y(\theta_{ML}) \\
&- \beta_k \nabla^2_{01} I_y(\theta_{ML}, \theta_{ML})\big)(\theta^{k+1} - \theta_{ML})\|.
\end{aligned} \tag{4.25}$$

Now, consider again the bounded set $\mathcal{B}$ containing $\{\theta^k\}$. Let λ_{l_y} and λ_I denote the minima

$$\lambda_{l_y} = \min_{\theta \in \mathcal{B}} \left\{-\lambda_{\nabla^2 l_y(\theta)}\right\}$$

$$\lambda_I = \min_{\theta, \bar\theta \in \mathcal{B}} \left\{\lambda_{\nabla^2_{01} I_y(\bar\theta, \theta)}\right\}.$$

Since for any symmetric matrix H, $x^T H x / \|x\|^2$ is lower bounded by the minimum eigenvalue of H, we have immediately that

$$\begin{aligned}
&\|\big(-\nabla^2 l_y(\theta_{ML}) + \beta_k \nabla^2_{01} I_y(\theta_{ML}, \theta_{ML})\big)(\theta^{k+1} - \theta_{ML})\|^2 \\
&\geqslant \big(\lambda_{l_y} + \beta_k \lambda_I\big)^2 \|\theta^{k+1} - \theta_{ML}\|^2.
\end{aligned} \tag{4.26}$$

By Assumptions 5.1.iii and 5.1.iv, $\lambda_{l_y} + \beta_k \lambda_I > 0$ and, after substitution of (4.26) into (4.25), we obtain

$$\begin{aligned}
\beta_k L\|\theta^k - \theta_{ML}\| + \|R(\theta^{k+1} - \theta_{ML})\| \geqslant \\
\big(\lambda_{l_y} + \beta_k \lambda_I\big)\|\theta^{k+1} - \theta_{ML}\|,
\end{aligned} \tag{4.27}$$

for all $\theta \in \mathcal{B}$. Therefore, collecting terms in (4.27)

$$\beta_k L \geqslant \left(\lambda_{l_y} + \beta_k \lambda_I - \frac{\|R(\theta^{k+1} - \theta_{ML})\|}{\|\theta^{k+1} - \theta_{ML}\|}\right) \frac{\|\theta^{k+1} - \theta_{ML}\|}{\|\theta^k - \theta_{ML}\|}. \tag{4.28}$$

Now, recall that $\{\theta^k\}$ is convergent. Thus, $\lim_{k\to\infty} \|\theta^k - \theta_{ML}\| = 0$ and subsequently, $\lim_{k\to\infty} \frac{\|R(\theta^{k+1} - \theta_{ML})\|}{\|\theta^{k+1} - \theta_{ML}\|} = 0$ due to the definition of the remainder R. Finally, as β_k converges to zero, L is bounded and $\lambda_{l_y} > 0$, equation (4.28) gives (4.20) with $\theta^* = \theta_{ML}$ and the proof of superlinear convergence is completed.

5 Second order Approximations and Trust Region techniques

The maximization in the KPP recursion (3.12) will not generally yield an explicit exact recursion in θ^k and θ^{k+1}. Thus implementation of the KPP algorithm methods may require line search or one-step-late approximations similar to those used for the M-step of the non-explicit penalized EM maximum likelihood algorithm [13]. In this section, we discuss an alternative which uses second order function approximations and preserves the convergence properties of KPP established in the previous section. This second order scheme is related to the well-known Trust Region technique for iterative optimization introduced by Moré [32].

Approximate models

In order to obtain computable iterations, the following second order approximations of $l_y(\theta)$ and $I_y(\theta^k, \theta)$ are introduced

$$\begin{aligned}\hat{l}_y(\theta) &= l_y(\theta^k) + \nabla l_y(\theta^k)^T(\theta - \theta^k) + \\ &\quad \frac{1}{2}(\theta - \theta^k)^T H_k(\theta - \theta^k).\end{aligned}$$

and

$$\hat{I}_y(\theta, \theta^k) = \frac{1}{2}(\theta - \theta^k)^T \nabla_{01}^2 I_k(\theta - \theta^k).$$

In the following, we adopt the simple notation $g_k = \nabla l_y(\theta^k)$ (a column vector). A natural choice for H_k and I_k is of course

$$H_k = \nabla^2 l_y(\theta^k)$$

and

$$I_k = \nabla_{01}^2 I_y(\theta^k, \theta^k).$$

The approximate KPP algorithm is defined as

$$\begin{aligned}\theta^{k+1} = \operatorname{argmax}_{\theta \in \mathbb{R}^p} \big\{ & l_y(\theta^k) + g_k(\theta - \theta^k) \\ & + \frac{1}{2}(\theta - \theta^k)^T H_k(\theta - \theta^k) \\ & - \frac{\beta_k}{2}(\theta - \theta^k)^T I_k(\theta - \theta^k) \big\}\end{aligned} \tag{5.29}$$

At this point it is important to make several comments. Notice first that for $\beta_k = 0$, $k = 1, 2, \ldots$, and $H_k = \nabla^2 l_y(\theta^k)$, the approximate step (5.29) is equivalent to a Newton step. It is well known that Newton's method, also known as Fisher scoring, has superlinear asymptotic convergence rate but may diverge if not properly initialized. Therefore, at least for small values of the relaxation parameter β_k, the approximate PPA algorithm may fail to converge for reasons analogous in Newton's method [37]. On the other hand, for $\beta_k > 0$ the term $-\frac{\beta_k}{2}(\theta - \theta^k)^T I_k(\theta - \theta^k)$ penalizes the distance of the next iterate θ^{k+1} to the current iterate θ^k. Hence, we can interpret this term as a regularization or relaxation which stabilizes the possibly divergent Newton algorithm without sacrificing its superlinear asymptotic convergence rate. By appropriate choice of $\{\beta_k\}$ the iterate θ^{k+1} can be forced to remain in a region around θ^k over which the quadratic model $\hat{l}_y(\theta)$ is accurate [32][7].

In many cases a quadratic approximation of a single one of the two terms $l_y(\theta)$ or $I_y(\theta^k, \theta)$ is sufficient to obtain a closed form for the maximum in the KPP recursion (3.12). Naturally, when feasible, such a reduced approximation is preferable to the approximation

of both terms discussed above. For concreteness, in the sequel, although our results hold for the reduced approximation also, we only prove convergence for the proximal point algorithm implemented with the full two-term approximation.

Finally, note that (5.29) is quadratic in θ and the minimization problem clearly reduces to solving a linear system of equations. For θ of moderate dimension, these equations can be efficiently solved using conjugate gradient techniques [34]. However, when the vector θ in (5.29) is of large dimension, as frequently occurs in inverse problems, limited memory BFGS quasi-Newton schemes for updating $H_k - \beta_k I_k$ may be computationally much more efficient, see for example [34], [35], [27], [12] and [11].

Trust Region Update Strategy

The Trust Region strategy proceeds as follows. The model $\hat{l}_y(\theta)$ is maximized in a ball $B(\theta^k, \delta) = \{\|\theta - \theta^k\|_{I_k} \leqslant \delta\}$ centered at θ^k where δ is a proximity control parameter which may depend on k, and where $\|a\|_{I_k} = a^T I_k a$ is a norm; well defined due to positive definiteness of I_k (Assumption 5.1.iv). Given an iterate θ^k consider a candidate θ^δ for θ^{k+1} defined as the solution to the constrained optimization problem

$$\theta^\delta = \text{argmax}_{\theta \in \mathbb{R}^p} \hat{l}_y(\theta)$$

subject to

$$\|\theta - \theta^k\|_{I_k} \leqslant \delta. \tag{5.30}$$

By duality theory of constrained optimization [16], and the fact that $\hat{l}_y(\theta)$ is strictly concave, this problem is equivalent to the unconstrained optimization

$$\theta^\delta(\beta) = \text{argmin}_{\theta \in \mathbb{R}^p} L(\theta, \beta). \tag{5.31}$$

where

$$L(\theta, \beta) = -\hat{l}_y(\theta) + \frac{\beta}{2}\big(\|\theta - \theta^k\|^2_{I_k} - \delta^2\big).$$

and β is a Lagrange multiplier selected to meet the constraint (3) with equality: $\|\theta^\delta(\beta) - \theta\|_{I_k} = \delta$.

We conclude that the Trust Region candidate θ^δ is identical to the approximate KPP iterate (5.29) with relaxation parameter β chosen according to constraint (3). This relation also provides a rational rule for computing the relaxation parameter β.

Implementation

The parameter δ is said to be safe if θ^δ produces an acceptable increase in the original objective l_y. An iteration of the Trust Region method consists of two principal steps

Rule 1. Determine whether δ is safe or not. If δ is safe, set $\delta_k = \delta$ and take an approximate Kullback proximal step $\theta^{k+1} = \theta^\delta$. Otherwise, take a null step $\theta^{k+1} = \theta^k$.

Rule 2. Update δ depending on the result of Rule 1.

Rule 1 can be implemented by comparing the increase in the original log-likelihood l_y to a fraction m of the expected increase predicted by the approximate model $\hat{l}_y(\theta)$. Specifically, the Trust Region parameter δ is accepted if

$$l_y(\theta^\delta) - l_y(\theta^k) \geqslant m\big(\hat{l}_y(\theta^\delta) - \hat{l}_y(\theta^k)\big). \tag{5.32}$$

Rule 2 can be implemented as follows. If δ was accepted by Rule 1, δ is increased at the next iteration in order to extend the region of validity of the model $\hat{l}_y(\theta)$. If δ was rejected, the region must be tightened and δ is decreased at the next iteration.

The Trust Region strategy implemented here is essentially the same as that proposed by Moré [32].

Step 0. (Initialization) Set $\theta^0 \in \mathbb{R}^p$, $\delta_0 > 0$ and the "curve search" parameters m, m' with $0 < m < m' < 1$.
Step 1. With $\hat{l}_y(\theta)$ the quadratic approximation (5.29), solve

$$\theta^{\delta_k} = \text{argmax}_{\theta \in \mathbb{R}^p} \hat{l}_y(\theta)$$

subject to

$$\|\theta - \theta^k\|_{I_k} \leqslant \delta_k.$$

Step 2. If $l_y(\theta^{\delta_k}) - l_y(\theta^k) \geqslant m\big(\hat{l}_y(\theta^{\delta_k}) - \hat{l}_y(\theta^k)\big)$ then set $\theta^{k+1} = \theta^{\delta_k}$. Otherwise, set $\theta^{k+1} = \theta^k$.
Step 3. Set $k = k + 1$. Update the model $\hat{l}_y(\theta^k)$. Update δ_k using Procedure 5.
Step 4. Go to Step 1.

The procedure for updating δ_k is given below.

Step 0. (Initialization) Set γ_1 and γ_2 such that $\gamma_1 < 1 < \gamma_2$.
Step 1. If $l_y(\theta^{\delta_k}) - l_y(\theta^k) \leqslant m\big(\hat{l}_y(\theta^{\delta_k}) - \hat{l}_y(\theta^k)\big)$ then take $\delta_{k+1} \in (0, \gamma_1 \delta_k)$.
Step 2. If $l_y(\theta^{\delta_k}) - l_y(\theta^k) \leqslant m'\big(\hat{l}_y(\theta^{\delta_k}) - \hat{l}_y(\theta^k)\big)$ then take $\delta_{k+1} \in (\gamma_1 \delta_k, \delta_k)$.
Step 3. If $l_y(\theta^{\delta_k}) - l_y(\theta^k) \geqslant m'\big(\hat{l}_y(\theta^{\delta_k}) - \hat{l}_y(\theta^k)\big)$ then take $\delta_{k+1} \in (\delta_k, \gamma_2 \delta_k)$.

The Trust Region algorithm satisfies the following convergence theorem

Theorem 5.1 *Let $g(y;\theta)$ and $k(x|y;\theta)$ be such that Assumptions 1 are satisfied. Then, $\{\theta^k\}$ generated by Algorithm 5 converges to the maximizer θ_{ML} of the log-likelihood $l_y(\theta)$ and satisfies the monotone likelihood property $l_y(\theta^{k+1}) \geqslant l_y(\theta^k)$. If in addition, the sequence of Lagrange multipliers $\{\beta_k\}$ tends towards zero, $\{\theta^k\}$ converges superlinearly.*

The proof of Theorem 5.1 is omitted since it is standard in the analysis of Trust Region methods; see [32, 34]. Superlinear convergence for the case that $\lim_{k\to\infty} \beta_k = 0$ follows from the Dennis and Moré criterion [7, Theorem 3.11].

Discussion

The convergence results of Theorems 1 and 2 apply to any class of objective functions which satisfy the Assumptions 5.1. For instance, the analysis directly applies to the penalized maximum likelihood (or posterior likelihood) objective function $l_y'(\theta) = l_y(\theta) + p(\theta)$ when the ML penalty function (prior) $p(\theta)$ is quadratic and non-negative of the form $p(\theta) = (\theta - \theta_o)^T R(\theta - \theta_o)$, where R is a non-negative definite matrix.

The convergence Theorems 1 and 2 make use of concavity of $l_y(\theta)$ and convexity of $I_y(\bar{\theta}, \theta)$ via Assumptions 5.1.iii and 5.1.iv. However, for smooth non-convex functions an analogous local superlinear convergence result can be established under somewhat stronger assumptions similar to those used in [15]. Likewise the Trust Region framework can also be applied to nonconvex objective functions. In this case, global convergence to a local maximizer of $l_y(\theta)$ can be established under Assumptions 5.1.i, 5.1.ii and 5.1.iv following the proof technique of [32].

6 Application to Poisson data

In this section, we illustrate the application of Algorithm 5 for a maximum likelihood estimation problem in a Poisson inverse problem arising in radiography, thermionic emission processes, photo-detection, and positron emission tomography (PET).

The Poisson Inverse Problem

The objective is to estimate the intensity vector $\theta = [\theta_1, \ldots, \theta_p]^T$ governing the number of gamma-ray emissions $N = [N_1, \ldots, N_p]^T$ over an imaging volume of p pixels. The estimate of θ must be based on a vector of m observed projections of N denoted $Y = [Y_1, \ldots, Y_m]^T$. The components N_i of N are independent Poisson distributed with rate parameters θ_i, and the components Y_j of Y are independent Poisson distributed with rate parameters $\sum_{i=1}^{p} P_{ji}\theta_i$, where P_{ji} is the transition probability; the probability that an emission from pixel i is detected at detector module j. The standard choice of complete data X, introduced by Shepp and Vardi [39], for the EM algorithm is the set $\{N_{ji}\}_{1 \leqslant j \leqslant m,\ 1 \leqslant i \leqslant p}$, where N_{ji} denotes the number of emissions in pixel i which are detected at detector j. The corresponding many-to-one mapping $h(X) = Y$ in the EM algorithm is

$$Y_j = \sum_{i=1}^{p} N_{ji}, \quad 1 \leqslant j \leqslant m. \tag{6.33}$$

It is also well known [39] that the likelihood function is given by

$$\log g(y;\theta) = \sum_{j=1}^{m} \Big(\sum_{i=1}^{p} P_{ji}\theta_i \Big) - y_j \log \Big(\sum_{i=1}^{p} P_{ji}\theta_i \Big) + \log y_j! \tag{6.34}$$

and that the expectation step of the EM algorithm is (see [13])

$$Q(\theta, \bar{\theta}) = \mathbb{E}[\log f(x;\theta) \mid y; \bar{\theta}] = \tag{6.35}$$

$$\sum_{j=1}^{m} \sum_{i=1}^{p} \Big(\frac{y_j P_{ji} \bar{\theta}_i}{\sum_{i=1}^{p} P_{ji} \bar{\theta}_i} \log(P_{ji}\theta_i) - P_{ji}\theta_i \Big).$$

Let us make the following additional assumptions:

- the solution(s) of the Poisson inverse problem is (are) positive
- the level set

$$\mathcal{L} = \{\theta \in \mathbb{R}^n \mid l_y(\theta) \geqslant l_y(\theta^1)\} \tag{6.36}$$

  is bounded and included in the positive orthant.

Then, since l_y is continuous, $\mathcal{L}$ is compact. Due to the monotonicity property of $\{\theta^k\}$, we thus deduce that for all k, $\theta_i^k \geqslant \gamma$ for some $\gamma > 0$. Then, the likelihood function and the regularization function are both twice continuously differentiable on the closure of $\{\theta^k\}$ and the theory developed in this paper applies. These assumptions are very close in spirit to the assumptions in Hero and Fessler [15], except that we do not require the maximizer to be unique. The study of KPP without these assumptions requires further analysis and is addressed in [6].

Simulation results

For illustration we performed numerical optimization for a simple one dimensional deblurring example under the Poisson noise model of the previous section. This example easily generalizes to more general 2 and 3 dimensional Poisson deblurring, tomographic reconstruction, and other imaging applications. The true source θ is a two rail phantom shown in Figure B.1. The blurring kernel is a Gaussian function yielding the blurred phantom shown in Figure A.2. We implemented both EM and KPP with Trust Region update strategy for deblurring Fig. A.2 when the set of ideal blurred data $Y_i = \sum_{j=1}^{N} P_{ij}\theta_j$ is available without Poisson noise. In this simple noiseless case the ML solution is equal to the true source θ which is everywhere positive. Treatment of this noiseless case allows us to investigate the behavior of the algorithms in the asymptotic high count rate regime. More extensive simulations with Poisson noise will be presented elsewhere.

The numerical results shown in Fig. B.2 indicate that the Trust Region implementation of the KPP algorithm enjoys significantly faster convergence towards the optimum than does EM. For these simulations the Trust Region technique was implemented in the standard manner where the trust region size sequence δ_k in Algorithm 1 is determined implicitly by the β_k update rule: $\beta_{k+1} = 1.6\beta_k$ (δ_k is decreased) and otherwise $\beta_{k+1} = 0.5\beta_k$ (δ_k is increased). The results shown in Fig. B.3 validate the theoretical superlinear convergence of the Trust Region iterates as contrasted with the linear convergence rate of the EM iterates. Figure **??** shows the reconstructed profile and demonstrates that the Trust Region updated KPP technique achieves better reconstruction of the original phantom for a fixed number of iterations. Finally, Figure **??** shows the iterates for the reconstructed phantom, plotted as a function of iteration on the horizontal axis and as a function of grey level on the vertical axis. Observe that the KPP achieves more rapid separation of the two components in the phantom than does standard EM.

7 Conclusions

The main contributions of this paper are the following. First, we introduced a general class of iterative methods for ML estimation based on Kullback-Liebler relaxation of the proximal point strategy. Next, we proved that the EM algorithm belongs to the proposed class, thus providing a new and useful interpretation of the EM approach for ML estimation. Finally, we showed that Kullback proximal point methods enjoy global convergence and even superlinear convergence for sequences of positive relaxation parameters that converge to zero. Implementation issues were also discussed and we proposed second order schemes for the case where the maximization step is hard to obtain in closed form. We addressed Trust Region methodologies for the updating of the relaxation parameters. Computational experiments indicated that the approximate second order KPP is stable and verifies the superlinear convergence property as was predicted by our analysis.

Figure A.1: Two rail phantom for 1D deblurring example.

Figure A.2: Blurred two level phantom. Blurring kernel is Gaussian with standard width approximately equal to rail separation distance in phantom. An additive randoms noise of 0.3 was added.

Figure A.3: Snapshot of log–Likelihood vs iteration for plain EM and KPP EM algorithm. Plain EM initially produces greater increases in likelihood function but is overtaken by KPP EM at 7 iterations and thereafter.

Figure A.4: The sequence $\log \|\theta_k - \theta^*\|$ vs iteration for plain EM and KPP EM algorithms. Here θ^* is limiting value for each of the algorithms. Note the superlinear convergence of KPP.

Bibliography

[1] A. A. Goldstein and I. B. Russak, "How good are the proximal point algorithms?," *Numer. Funct. Anal. and Optimiz.*, vol. 9, no. 7-8, pp. 709–724, 1987. (p. 44).

[2] H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Review*, vol. 38, no. 3, pp. 367–426, 1996. (p. 50).

[3] J. F. Bonnans, J.-C. Gilbert, C. Lemaréchal, and C. Sagastizabal, *Optimization numérique. Aspects théoriques et pratiques*, volume 27, Springer Verlag, 1997. Series : Mathématiques et Applications. (pp. 53 et 55).

[4] C. Bouman and K. Sauer, "Fast numerical methods for emission and transmission tomographic reconstruction," in *Proc. Conf. on Inform. Sciences and Systems*, Johns Hopkins, 1993. (p. 44).

[5] Y. Censor and S. A. Zenios, "Proximal minimization algorithm with D-functions," *Journ. Optim. Theory and Appl.*, vol. 73, no. 3, pp. 451–464, June 1992. (p. 44).

[6] S. Chrétien and A. Hero, "Generalized proximal point algorithms," *SIAM Journ. on Optimization*, Submitted Sept., 1998. (p. 56).

[7] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, 1987. (p. 49).

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society, Ser. B*, vol. 39, no. 1, pp. 1–38, 1977. (pp. 43, 45, 46 et 86).

[9] P. Eggermont, "Multiplicative iterative algorithms for convex programming.," *Linear Algebra Appl.*, vol. 130, pp. 25–42, 1990. (p. 44).

[10] J. Ekstein, "Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming," *Math. Oper. Res.*, vol. 18, no. 1, pp. 203–226, February 1993. (p. 44).

[11] R. Fletcher, "A new variational result for quasi-Newton formulae.," *SIAM J. Optim.*, vol. 1, no. 1, pp. 18–21, 1991. (p. 54).

[12] J. C. Gilbert and C. Lemarechal, "Some numerical experiments with variable-storage quasi-Newton algorithms.," *Math. Program., Ser. B*, vol. 45, no. 3, pp. 407–435, 1989. (p. 54).

[13] P. J. Green, "On the use of the EM algorithm for penalized likelihood estimation," *J. Royal Statistical Society, Ser. B*, vol. 52, no. 2, pp. 443–452, 1990. (pp. 53 et 56).

[14] D. Helmbold, R. Schapire, S. Y., and W. M., "A comparison of new and old algorithms for a mixture estimation problem," *Journal of Machine Learning*, vol. 27, no. 1, pp. 97–119, 1997. (p. 45).

[15] A. O. Hero and J. A. Fessler, "Convergence in norm for alternating expectation-maximization (EM) type algorithms," *Statistica Sinica*, vol. 5, no. 1, pp. 41–54, 1995. (pp. 43, 55 et 56).

[16] J. B. Hiriart-Hurruty and C. Lemaréchal, *Convex analysis and minimization algorithms I-II*, Springer-Verlag, Bonn, 1993. (pp. 44, 47 et 54).

[17] H. Hudson and R. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE Transactions on Medical Imaging*, vol. 13, no. 12, pp. 601–609, 1994. (p. 44).

[18] I. A. Ibragimov and R. Z. Has'minskii, *Statistical estimation: Asymptotic theory*, Springer-Verlag, New York, 1981. (pp. 47 et 89).

[19] M. Jamshidian and R. I. Jennrich, "Conjugate gradient acceleration of the EM algorithm," *J. Am. Statist. Assoc.*, vol. 88, no. 421, pp. 221–228, 1993. (p. 44).

[20] L. Kaufman, "Implementing and accelerating the EM algorithm for positron emission tomography," *IEEE Trans. on Medical Imaging*, vol. MI-6, no. 1, pp. 37–51, 1987. (p. 44).

[21] J. Kivinen and M. K. Warmuth, "Additive versus exponentiated gradient updates for linear prediction," *Information and Computation*, vol. 132, pp. 1–64, January 1997. (p. 45).

[22] K. Lange, "A quasi-newtonian acceleration of the EM algorithm," *Statistica Sinica*, vol. 5, no. 1, pp. 1–18, 1995. (p. 44).

[23] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *Journal of Computer Assisted Tomography*, vol. 8, no. 2, pp. 306–316, April 1984. (p. 43).

[24] D. Lansky and G. Casella, "Improving the EM algorithm," in *Computing and Statistics: Proc. Symp. on the Interface*, C. Page and R. LePage, editors, pp. 420–424, Springer-Verlag, 1990. (p. 44).

[25] M. Lavielle, "Stochastic algorithm for parametric and non-parametric estimation in the case of incomplete data," *Signal Processing*, vol. 42, no. 1, pp. 3–17, 1995. (p. 44).

[26] R. Lewitt and G. Muehllehner, "Accelerated iterative reconstruction for positron emission tomography," *IEEE Trans. on Medical Imaging*, vol. MI-5, no. 1, pp. 16–22, 1986. (p. 44).

[27] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization.," *Math. Program., Ser. B*, vol. 45, no. 3, pp. 503–528, 1989. (p. 54).

[28] T. A. Louis, "Finding the observed information matrix when using the EM algorithm," *J. Royal Statistical Society, Ser. B*, vol. 44, no. 2, pp. 226–233, 1982. (p. 44).

[29] B. Martinet, "Régularisation d'inéquation variationnelles par approximations successives," *Revue Francaise d'Informatique et de Recherche Operationnelle*, vol. 3, pp. 154–179, 1970. (p. 44).

[30] I. Meilijson, "A fast improvement to the EM algorithm on its own terms," *J. Royal Statistical Society, Ser. B*, vol. 51, no. 1, pp. 127–138, 1989. (p. 44).

[31] G. J. Minty, "Monotone (nonlinear) operators in Hilbert space," *Duke Math. Journal*, vol. 29, pp. 341–346, 1962. (p. 44).

[32] J. J. Moré, "Recent developments in algorithms and software for trust region methods," in *Mathematical programming: the state of the art*, pp. 258–287, Springer-Verlag, Bonn, 1983. (pp. 45, 53 et 55).

[33] J. J. Moreau, "Proximité et dualité dans un espace Hilbertien," *Bull. Soc. Math. France*, vol. 93, pp. 273–299, 1965. (p. 44).

[34] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Series in Operations Research, Springer Verlag, Berlin, 1999. (pp. 45, 54 et 55).

[35] J. Nocedal, "Updating quasi-Newton matrices with limited storage.," *Math. Comput.*, vol. 35, pp. 773–782, 1980. (p. 54).

[36] J. M. Ollinger and D. L. Snyder, "A preliminary evaluation of the use of the EM algorithm for estimating parameters in dynamic tracer studies," *IEEE Trans. Nuclear Science*, vol. NS-32, pp. 3575–3583, Feb. 1985. (p. 43).

[37] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970. (p. 53).

[38] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization*, vol. 14, pp. 877–898, 1976. (pp. 44, 47, 63 et 67).

[39] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. on Medical Imaging*, vol. MI-1, No. 2, pp. 113–122, Oct. 1982. (pp. 43, 44 et 56).

[40] M. Teboulle, "Entropic proximal mappings with application to nonlinear programming," *Mathematics of Operations Research*, vol. 17, pp. 670–690, 1992. (pp. 44, 67 et 96).

[41] M. Warmuth and K. Azoury, "Relative loss bounds for on-line density estimation with the exponential family of distributions," *Proc. of Uncertainty in Artif. Intel.*, 1999. (p. 45).

[42] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Annals of Statistics*, vol. 11, pp. 95–103, 1983. (p. 43).

Chapter B

A Component-wise EM Algorithm for Mixtures

with Gilles Celeux, Florence Forbes and Abdallah Mkhadri.

Abstract

In some situations, EM algorithm shows slow convergence problems. One possible reason is that standard procedures update the parameters simultaneously. In this paper we focus on finite mixture estimation. In this framework, we propose a component-wise EM, which updates the parameters sequentially. We give an interpretation of this procedure as a proximal point algorithm and use it to prove the convergence. Illustrative numerical experiments show how our algorithm compares to EM and a version of the SAGE algorithm.

1 Introduction

Estimation in finite mixture distributions is typically an incomplete data structure problem for which the EM algorithm [3] is used (see for instance [4]). The most documented problem occuring with the EM algorithm is its possible low speed in some situations. Many papers, including [22], [18], [19], [21], [20] have proposed extensions of the EM algorithm based on standard numerical tools to speed up the convergence. There are often effective, but they do not guarantee monotone increase in the objective function. To overcome this problem, alternatives based on model reduction ([41],[42]) and efficient data augmentation ([17], [16], [15], [23], [24], [35], [43], see also the chapter 5 of [32]) have recently been considered. These extensions share the simplicity and stability with EM while speeding up the convergence. However, as far as we know, only two extensions ([44], [40]) were devoted to speeding up the convergence in the mixture case which is one of the most important domains of application for EM. The first one [44] is based on a restricted efficient data augmentation scheme for the estimation of the proportions for known discrete distributions. While the second extension [40] is concerned with the implementation of the ECME algorithm ([42]) for mixture distributions.

In this paper we propose, study and illustrate a component-wise EM algorithm (CEM^2: Component-wise EM algorithm for Mixtures) aiming at overcoming the slow convergence problem in the finite mixture context. Our approach is based on a recent work [33], [34], [2] which recasts the EM procedure in the framework of proximal point algorithms [9] and

[12]. In Section 2 we present the EM algorithm for mixtures and its interpretation as a proximal point algorithm. In Section 3, we describe our component-wise algorithm and show, in Section 4, that it can also be interpreted as a proximal point algorithm. Using this interpretation, convergence of CEM^2 is proved in Section 5. Illustrative numerical experiments comparing the behaviors of EM, a version of the SAGE algorithm [17, 16] and CEM^2 are presented in Section 6. A discussion section ends the paper. An appendix carefully describes the SAGE method in the mixture context in order to provide detailed comparison with the proposed CEM^2.

2 EM-type algorithms for mixtures

We consider a J-component mixture in $\mathbb{R}^d$

$$g(y|\theta) = \sum_{j=1}^{J} p_j \varphi(y|\alpha_j) \tag{2.1}$$

where the p_j's ($0 < p_j < 1$ and $\sum j = 1 J p_j = 1$) are the mixing proportions and where $\varphi(y|\alpha)$ is a density function parametrized by α. The vector parameter to be estimated is $\theta = (p_1, \ldots, p_J, \alpha_1, \ldots, \alpha_J)$.

The parametric families of mixture densities are assumed to be identifiable. This means that for any two members of the form (2),

$$g(y|\theta) \equiv g(y|\theta')$$

if and only if $J = J'$ and we can permute the components labels so that $p_j = p_{j'}$ and $\varphi(y|\alpha_j) = \varphi(y|\alpha_{j'})$, for $j = 1, \ldots, J$. Most mixtures of interest are identifiable (see for instance [4]).

For the sake of simplicity, we restrict the present analysis to Gaussian mixtures, but extension to more general mixtures is straightforward (as long as the considered densities are differentiable functions of the paramater α). Thus, $\varphi(y|\mu, \Sigma)$ denotes the density of a Gaussian distribution with mean μ and variance matrix Σ. The parameter to be estimated is

$$\theta = (p_1, \ldots, p_J, \mu_1, \ldots, \mu_J, \Sigma_1, \ldots, \Sigma_J).$$

In the following, we denote $\theta_j = (p_j, \mu_j, \Sigma_j)$, for $j = 1, \ldots, J$.

The EM algorithm

The mixture density estimation problem is typically a missing data problem for which the EM algorithm appears to be useful.

Let $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^{dn}$ be an observed sample from the mixture distribution $g(y|\theta)$. We assume that the component from which each y_i arised is unknown so that the missing data are the labels z_i $i = 1, \ldots, n$. We have $z_i = j$ if and only if j is the mixture component from which y_i arises. Let $\mathbf{z} = (z_1, \ldots, z_n)$ denote the missing data, $\mathbf{z} \in B^n$, where $B = \{1, \ldots, J\}$. The complete sample is $\mathbf{x} = (x_1, \ldots, x_n)$ with $x_i = (y_i, z_i)$. We have $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ and the non-invertible transformation π such that $\mathbf{y} = \pi(\mathbf{x})$ is the projection of $\mathbb{R}^{dn} \times B^n$ on $\mathbb{R}^{dn}$. The observed log-likelihood is

$$L(\theta|\mathbf{y}) = \log \mathbf{g}(\mathbf{y}|\theta),$$

where $\mathbf{g}(\mathbf{y}|\theta)$ denotes the density of the observed sample $\mathbf{y}$. Using (2) leads to

$$L(\theta|\mathbf{y}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{J} p_j \varphi(y_i|\mu_j, \Sigma_j) \right\}.$$

The complete log-likelihood is

$$L(\theta|\mathbf{x}) = \log \mathbf{f}(\mathbf{x}|\theta),$$

where $\mathbf{f}(\mathbf{x}|\theta)$ denotes the density of the complete sample $\mathbf{x}$. We have

$$L(\theta|\mathbf{x}) = \sum_{i=1}^{n} \{\log p_{z_i} + \log \varphi(y_i|\mu_{z_i}, \Sigma_{z_i})\}. \tag{2.2}$$

The conditional density function of the complete data given $\mathbf{y}$

$$\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta) = \frac{\mathbf{f}(\mathbf{x}|\theta)}{\mathbf{g}(\mathbf{y}|\theta)} \tag{2.3}$$

takes the form

$$\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta) = \prod_{i=1}^{n} t_{iz_i}(\theta) \tag{2.4}$$

where $t_{ij}(\theta), j = 1, \ldots, J$ denotes the conditional probability, given $\mathbf{y}$, that y_i arises from the mixture component with density $\varphi(.|\mu_j, \Sigma_j)$. ¿From Bayes formula, we have for each i $(1 \leqslant i \leqslant n)$ and j $(1 \leqslant j \leqslant J)$

$$t_{ij}(\theta) = \frac{p_j \varphi(y_i|\mu_j, \Sigma_j)}{\sum_{\ell=1}^{J} p_\ell \varphi(y_i|\mu_\ell, \Sigma_\ell)}. \tag{2.5}$$

Thus the conditional expectation of the complete log-likelihood given $\mathbf{y}$ and a previous estimate of θ, denoted θ',

$$Q(\theta|\theta') = I\!E\left[\log L(\theta|\mathbf{x})|\mathbf{y}, \theta'\right]$$

takes the form

$$Q(\theta|\theta') = \sum_{i=1}^{n} \sum_{\ell=1}^{J} t_{i\ell}(\theta') \{\log p_\ell + \log \varphi(y_i|\mu_\ell, \Sigma_\ell)\}. \tag{2.6}$$

The EM algorithm generates a sequence of approximations to find the maximum observed likelihood estimator starting from an initial guess θ^0, using two steps. The kth iteration is as follows

E-step: Compute $Q(\theta|\theta^k) = I\!E\left[\log \mathbf{f}(\mathbf{x}|\theta)|\mathbf{y}, \theta^k\right]$.
M-step: Find $\theta^{k+1} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^k)$, where $\Theta = \{(p_1, \ldots, p_J, \alpha_1, \ldots, \alpha_J)\}$.

In many situations, including the mixture case, the explicit computation of $Q(\theta|\theta^k)$ in the E-step is unnecessary and this step reduces to the computation of the conditional density $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta^k)$.

For Gaussian mixtures, these two steps take the form

E-step: For $i = 1, \ldots, n$ and $j = 1, \ldots, J$ compute

$$t_{ij}(\theta^k) = \frac{p_j^k \varphi(y_i|\mu_j^k, \Sigma_j^k)}{\sum_{\ell=1}^{J} p_\ell^k \varphi(y_i|\mu_\ell^k, \Sigma_\ell^k)}. \tag{2.7}$$

M-step : Set $\theta^{k+1} = (p_1^{k+1}, \ldots, p_J^{k+1}, \mu_1^{k+1}, \ldots, \mu_J^{k+1}, \Sigma_1^{k+1}, \ldots, \Sigma_J^{k+1})$ with

$$p_j^{k+1} = \frac{1}{n} \sum_{i=1}^{n} t_{ij}(\theta^k)$$

$$\mu_j^{k+1} = \frac{\sum i = 1 n t_{ij}(\theta^k)\, y_i}{\sum i = 1 n t_{ij}(\theta^k)} \tag{2.8}$$

$$\Sigma_j^{k+1} = \frac{\sum i = 1 n t_{ij}(\theta^k)(y_i - \mu_j^{k+1})(y_i - \mu_j^{k+1})^T}{\sum i = 1 n t_{ij}(\theta^k)}.$$

Note that at each iteration, the following properties hold

$$\text{for } i = 1, \ldots, n, \quad \sum_{j=1}^{J} t_{ij}(\theta^k) = 1$$

$$\text{and } \sum_{j=1}^{J} p_j^k = 1. \tag{2.9}$$

Proximal interpretation of the EM algorithm

The EM algorithm can be viewed as an alternating optimisation algorithm (see [35], or [25] in the mixture context). The function to be maximized takes the form

$$F(\mathbf{p}, \theta) = \int L(\theta|\mathbf{x})\mathbf{p}(\mathbf{z})d\mathbf{z} + H(\mathbf{p}), \tag{2.10}$$

where

$$H(\mathbf{p}) = -\int \mathbf{p}(\mathbf{z}) \log \mathbf{p}(\mathbf{z}) d\mathbf{z}$$

is the entropy of the probability distribution $\mathbf{p}$ defined on the missing data set which is B^n in the mixture context. Denoting $\mathcal{T}$ the set of probability distributions on the missing data set, an iteration of EM can be expressed as follows:

E-step: $\mathbf{t}^{k+1} = \arg\max_{\mathbf{t} \in \mathcal{T}} F(\mathbf{t}, \theta^k)$.
M-step: $\theta^{k+1} = \arg\max_{\theta \in \Theta} F(\mathbf{t}^{k+1}, \theta)$.

Here we detail a presentation of EM as a proximal point algorithm with a Kullback-Leibler-type penalty which includes the interpretation of EM as an alternating optimisation algorithm. Consider the general problem of maximizing a concave function $\Phi(\theta)$. Then, the proximal point algorithm is an iterative procedure which is defined by the following recurrence,

$$\theta^{k+1} = \arg\max_{\theta \in \mathbb{R}^p} \left\{ \Phi(\theta) - \frac{1}{2} \|\theta - \theta^k\|^2 \right\}. \tag{2.11}$$

In other words, the objective function Φ is regularized using a quadratic penalty $\|\theta - \theta^k\|^2$. The function

$$Y(\bar{\theta}) = \max_{\theta \in \mathbb{R}^p} \left\{ \Phi(\theta) - \frac{1}{2}\|\theta - \bar{\theta}\|^2 \right\} \tag{2.12}$$

is often called the Moreau-Yosida regularization of Φ. The proximal point algorithm was first studied in [9]. The proximal methodology was then applied to many types of algorithms and is still in great effervescence (see [19, 12] for instance and the literature therein).

As shown in [33], the EM procedure can be recast into a proximal point framework. This point of view provides much insight into the algorithm convergence properties. In particular it has already been shown in [34] that convergence holds without differentiability assumptions, hence appropriately handling the case of Laplace distributions, and in [33] that superlinear convergence of the iterates could be obtained under twice differentiability assumptions, usually satisfied by most distributions in pratice except the Laplace law. In this paper, the proximal formulation of EM for mixture densities is also of great importance and appears to be an essential tool in the convergence proof of the CEM^2 algorithm presented in Section 3.

We first introduce an appropriate Kullback information measure. Assume that the family of parametrized conditional densities $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)$ with $\theta \in \Theta$ defined in (2.3) is regular in the sense of Ibragimov and Khas'minskij [30], in particular $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)\lambda(d\mathbf{x})$ and $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta')\lambda(d\mathbf{x})$ are absolutely continuous with respect to each other for any θ and θ' in Θ, $\lambda(d\mathbf{x})$ being the product of the Lebesgue measure and the counting measure on $\mathbb{R}^{nd} \times B^n$. Then the Radon-Nikodym derivative $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta')/\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)$ exists for all θ, θ' and the following Kullback-Leibler divergence between vectors $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta')$ and $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)$ is well defined,

$$I(\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta'), \mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)) = I\!E\big[\log \frac{\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta')}{\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)} | \mathbf{y}; \theta'\big]. \tag{2.13}$$

In addition, (2.4) can be used as a measure of distance D between θ and θ' by setting

$$D(\theta, \theta'|\mathbf{y}) = I(\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta'), \mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)) \; . \tag{2.14}$$

In [33], the following proposition is established.

Proposition 2.1 ***(Chrétien and Hero 1998)*** *The EM algorithm is a proximal point algorithm with Kullback-type penalty (2.14) of the form*

$$\theta^{k+1} = \arg\max_{\theta \in \Theta} \left\{ L(\theta \mid \boldsymbol{y}) - D(\theta, \theta^k|\boldsymbol{y}) \right\}. \tag{2.15}$$

Hence, the EM algorithm can be interpreted as a generalized proximal point procedure where the quadratic Moreau-Yosida regularization is replaced by a Kullback information measure between the two conditional densities $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta)$ and $\mathbf{t}(\mathbf{x}|\mathbf{y}, \theta^k)$.

Note that in the mixture case, using (2.4), it comes

$$\begin{aligned} D(\theta, \theta'|\mathbf{y}) &= \sum i = 1 n I(t_{i.}(\theta'), t_{i.}(\theta)) \\ &= \sum_{i=1}^{n} \sum_{\ell=1}^{J} t_{i\ell}(\theta') \log \left(\frac{t_{i\ell}(\theta')}{t_{i\ell}(\theta)} \right) \; , \end{aligned} \tag{2.16}$$

where $I(t_{i.}(\theta'), t_{i.}(\theta))$ is the Kullback-Leibler divergence between vectors $t_{i.}(\theta') = (t_{i1}(\theta'), \ldots, t_{iJ}(\theta'))$ and $t_{i.}(\theta) = (t_{i1}(\theta), \ldots, t_{iJ}(\theta))$, which can be viewed as probability measures on $\{1, \ldots, J\}$ and that we further assume to be strictly positive. The $t_{i\ell}(\theta)$'s are defined in (2.5). Let $t(\theta)$

be the $n \times J$ probability matrix with general term $t_{i\ell}(\theta)$. A question of importance here is whether or not the following property holds,

$$D(\theta, \theta' | \mathbf{y}) = 0 \quad \Rightarrow \quad \theta' = \theta \; . \tag{2.17}$$

This is not generally the case when θ lies in $\mathbb{R}^p$ since $t(.)$ is not injective. Indeed, for θ and θ' in $\mathbb{R}^p$ such that, for $j = 1, \ldots, J$,

$$\begin{array}{rcl} p_j & = & \alpha \, p'_j \\ \mu_j & = & \mu'_j \\ \Sigma_j & = & \Sigma'_j \end{array}$$

for some $\alpha > 0$ different from one, it comes $t(\theta) = t(\theta')$ although $\theta \neq \theta'$. However, (2.17) holds when the constraint $\sum \ell = 1 J p_\ell = 1$ is satisfied. Then,

$$\begin{array}{rcl} Q(\theta \mid \theta^k) & = & \sum i = 1 n \sum \ell = 1 J t_{i\ell}(\theta^k) \log \dfrac{p_\ell \varphi(y_i \mid \mu_\ell, \Sigma_\ell)}{t_{i\ell}(\theta)} \\ & & + \sum i = 1 n \sum \ell = 1 J t_{i\ell}(\theta^k) \log t_{i\ell}(\theta) \\ & & - \sum i = 1 n \sum \ell = 1 J t_{i\ell}(\theta^k) \log t_{i\ell}(\theta^k) \\ & & + \sum i = 1 n \sum \ell = 1 J t_{i\ell}(\theta^k) \log t_{i\ell}(\theta^k) \; . \end{array}$$

Using (2.5), we can further write

$$\begin{array}{rcl} Q(\theta \mid \theta^k) & = & \sum i = 1 n \log \sum \ell = 1 J \Big\{ p_\ell \varphi(y_i \mid \mu_\ell, \Sigma_\ell) \Big\} \sum \ell = 1 J t_{i\ell}(\theta^k) \\ & & - D(\theta^k, \theta \mid \mathbf{y}) \\ & & + \sum i = 1 n \sum \ell = 1 J t_{i\ell}(\theta^k) \log t_{i\ell}(\theta^k) \; . \end{array}$$

Since $\sum \ell = 1 J t_{i\ell}(\theta^k) = 1$, it comes

$$Q(\theta \mid \theta^k) = L(\theta \mid \mathbf{y}) - D(\theta, \theta^k \mid \mathbf{y}) + \sum i = 1 n \sum \ell = 1 J t_{i\ell}(\theta^k) \log t_{i\ell}(\theta^k) \; . \tag{2.18}$$

The last term in the right-hand side does not depend on θ.

3 A Component-wise EM for mixtures

Component-wise methods have been introduced early in the computational literature. Serial decomposition of optimization methods is a well known procedure in numerical analysis. Assuming that θ lies in $\mathbb{R}^p$, the optimization problem

$$\max_{\theta \in \mathbb{R}^p} \Phi(\theta)$$

is decomposed into a series of coordinate-wise maximization problems of the form

$$\max_{\eta \in \mathbb{R}} \Phi(\theta_1, \ldots, \theta_{j-1}, \eta, \theta_{j+1}, \ldots, \theta_p).$$

This procedure is called a Gauss-Seidel scheme. The study of this method is standard (see [1] for example). The proximal method and the Gauss-Seidel scheme can be merged, which leads to the following recursion,

$$\begin{cases} \theta_j^{k+1} = \arg\max_{\eta \in \mathbb{R}} \left\{ \Phi(\theta_1^k, \ldots, \theta_{j-1}^k, \eta, \theta_{j+1}^k, \ldots, \theta_p^k) + \frac{1}{2} \| \eta - \theta_i^k \|^2 \right\} \\ \theta_i^{k+1} = \theta_i^k, \;\; i \neq j. \end{cases} \tag{3.19}$$

Component-wise methods aim at avoiding slow convergence situations. An intuitive idea is that exploring the parameter space sequentially rather than simultaneously tends to prevent from getting trapped in difficult situations (e.g. near saddle points). One of the most promising general purpose extention of EM, going in this direction, is the Space-Alternating Generalized EM (SAGE) algorithm [17]. Improved convergence rates are reached by updating the parameters sequentially in small groups associated to small hidden data spaces rather than one large complete data space. The SAGE method is very general and flexible. In the Appendix, more details are given in the mixture context. More specifically, since the SAGE approach is closely related to the CEM^2 algorithm, we describe, for comparison purpose, a version of SAGE for Gaussian mixtures. This version is nearly a component-wise algorithm except that the mixing proportions need to be updated in the same iteration, which involves the whole complete data structure. For this reason, it may not be significantly faster than the standard EM algorithm. This points out the main interest of the component-wise EM algorithm that we propose for mixtures. No iteration needs the whole complete data space as hidden-data space. It is a full component-wise algorithm and can therefore be expected to converge faster in various situations.

Our Component-wise EM algorithm for Mixtures (CEM^2) considers the decomposition of the parameter vector $\theta = (\theta_j, j = 1, ..., J)$ with $\theta_j = (p_j, \mu_j, \Sigma_j)$. The idea is to update only one component at a time, letting the other parameters unchanged. The order according to which the components are visited may be arbitrary, prescribed or varying adaptively. For simplicity, in our presentation, the components are updated successively, starting from $j = 1, \ldots, J$ and repeating this after J iterations. Therefore the component updated at iteration k is given by (3.20) and the kth iteration of the algorithm is as follows. For

$$j = k - \frac{k}{J}\rfloor J + 1, \tag{3.20}$$

$.\rfloor$ denoting the integer part, it alternates the following steps
E-step: Compute for $i = 1, \ldots, n$,

$$t_{ij}(\theta^k) = \frac{p_j^k \varphi(y_i|\mu_j^k, \Sigma_j^k)}{\sum_{\ell=1}^{J} p_\ell^k \varphi(y_i|\mu_\ell^k, \Sigma_\ell^k)}. \tag{3.21}$$

M-step: Set

$$\begin{aligned} p_j^{k+1} &= \frac{1}{n}\sum_{i=1}^{n} t_{ij}(\theta^k) \\ \mu_j^{k+1} &= \frac{\sum i = 1 n t_{ij}(\theta^k) y_i}{\sum i = 1 n t_{ij}(\theta^k)} \\ \Sigma_j^{k+1} &= \frac{\sum i = 1 n t_{ij}(\theta^k)(y_i - \mu_j^{k+1})(y_i - \mu_j^{k+1})^T}{\sum i = 1 n t_{ij}(\theta^k)}, \end{aligned} \tag{3.22}$$

and for $\ell \neq j$, $\theta_\ell^{k+1} = \theta_\ell^k$.

Note that the main difference with the SAGE algorithm presented in the Appendix is that the updating steps of the mixing proportions cannot be regarded as maximization steps of the form (7.55). Consequently, the SAGE standard assumptions are not satisfied and a specific convergence analysis must be achieved. It is based on the proximal interpretation of CEM^2 given in the next section.

4 Lagrangian and Proximal representation of CEM^2

Lagrangian approach

As underlined in the previous section, the main difficulty which prevents from passing to fully component-wise approaches resides in the treatment of the constraint $\sum \ell = 1 J p_\ell = 1$. This difficulty is usually dealt with by introduction of a reduced parameter space

$$\Omega = \left\{ \left(p_1, \ldots, p_{J-1}, \mu_1, \ldots, \mu_J, \Sigma_1, \ldots, \Sigma_J \right) \right\}, \tag{4.23}$$

the remaining proportion being trivially deduced from the $J-1$ others, knowing that

$$p_J = 1 - \sum \ell = 1 J - 1 p_\ell, \tag{4.24}$$

see [4] for instance. Obviously, this latter "reduced" representation of the parameter space is unsatisfactory in the context of coordinate-wise methods.

The linear constraint $\sum \ell = 1 J p_\ell = 1$ is easily handled via Lagrange duality in the following manner. Consider the *Lagrangian* function

$$\mathcal{L}(\theta, \lambda) = L(\theta|\mathbf{y}) - \lambda \Big(\sum \ell = 1 J p_\ell - 1 \Big). \tag{4.25}$$

The original constrained maximum likelihood problem can be reduced to the following unconstrained problem

$$\textbf{(primal)} \sup_{\theta \in \Theta} \inf_{\lambda \in \mathbb{R}} \mathcal{L}(\theta, \lambda), \tag{4.26}$$

where $\Theta = \{(p_1, \ldots, p_J, \mu_1, \ldots, \mu_J, \Sigma_1, \ldots, \Sigma_J)\}$. Indeed, when the constraints are not satisfied, the value in (4.26) is $-\infty$. Dualizing, we obtain the minimization problem

$$\textbf{(dual)} \inf_{\lambda \in \mathbb{R}} \sup_{\theta \in \Theta} \mathcal{L}(\theta, \lambda). \tag{4.27}$$

Although well known, the Lagrangian representation is rarely mentioned in the EM literature of mixture estimation. In this paper, the Lagrangian approach will give much insight in the proximal formulation below and thus, in the convergence proof of Section 5.

Generalized proximal point procedure

We first consider a Kullback proximal procedure in order to solve the maximum likelihood problem via the Lagrangian formulation. Then, we show that CEM^2 is a coordinate-wise maximization of the primal function in the Lagrangian framework.

The Kullback proximal regularization we consider is defined by

$$K(\bar{\theta}) = \sup_{\theta \in \Theta} L(\theta|\mathbf{y}) - D(\theta, \bar{\theta} \mid \mathbf{y}). \tag{4.28}$$

The proximal point iteration associated to this Kullback regularization is given by

$$\theta^{k+1} = \arg\max_{\theta \in \Theta} L(\theta|\mathbf{y}) - D(\theta, \theta^k \mid \mathbf{y}), \tag{4.29}$$

under the assumption that such a maximizer exists, which will be seen later as a natural assumption in the EM context. Applying this Kullback proximal iteration to the Lagrange representation (4.26) of the constrained maximum likelihood problem, we obtain

$$\theta^{k+1} = \arg\max_{\theta \in \Theta} \inf_{\lambda \in \mathbb{R}} L(\theta|\mathbf{y}) - D(\theta, \theta^k \mid \mathbf{y}) - \lambda \Big(\sum \ell = 1 J p_\ell - 1 \Big). \tag{4.30}$$

Now, dualizing as in (4.27), we obtain the new Kullback proximal iteration

$$(\lambda^{k+1}, \theta^{k+1}) = \arg\min_{\lambda\in\mathbb{R}} \arg\max_{\theta\in\Theta} L(\theta|\mathbf{y}) - D(\theta, \theta^k \mid \mathbf{y}) - \lambda\Big(\sum \ell = 1 J p_\ell - 1\Big) \tag{4.31}$$

under the assumption that the "argmin" exists, which will be shown below. Now, using Proposition 2.1, we obtain the following iteration

$$(\lambda^{k+1}, \theta^{k+1}) = \arg\min_{\lambda\in\mathbb{R}} \arg\max_{\theta\in\Theta} Q(\theta|\theta^k) - \lambda\Big(\sum \ell = 1 J p_\ell - 1\Big). \tag{4.32}$$

Define the function

$$\Delta(\lambda) = \max_{\theta\in\Theta} \left\{ Q(\theta|\theta^k) - \lambda\Big(\sum \ell = 1 J p_\ell - 1\Big) \right\}. \tag{4.33}$$

Replacing $Q(\theta|\theta^k)$ by its value deduced from (2.6), we obtain that, at the optimum in (4.33),

$$p_\ell = \sum_{i=1}^{n} t_{i\ell}(\theta^k)/\lambda \tag{4.34}$$

for all $\ell = 1, \ldots, J$. Therefore, we have

$$\Delta(\lambda) = \sum_{i=1}^{n}\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) \log \frac{\sum_{i=1}^{n} t_{i\ell}(\theta^k)}{\lambda} - \lambda\Big(\sum_{\ell=1}^{J} \frac{\sum_{i=1}^{n} t_{i\ell}(\theta^k)}{\lambda} - 1\Big) + \mathcal{R}, \tag{4.35}$$

where $\mathcal{R}$ is a remainder term independent of λ. Now, simple calculation gives the value of λ minimizing Δ,

$$\lambda^{k+1} = \sum_{i=1}^{n}\sum_{\ell=1}^{J} t_{i\ell}(\theta^k). \tag{4.36}$$

Since $t_{i\ell}(\theta^k)$ is a conditional probability, $\sum_{\ell=1}^{J} t_{i\ell}(\theta^k) = 1$. Thus, we obtain that

$$\lambda^{k+1} = n \tag{4.37}$$

for all k in $\mathbb{N}$. Finally, the Kullback proximal iteration applied to the Lagrangian dual becomes

$$\theta^{k+1} = \arg\max_{\theta\in\Theta} Q(\theta|\theta^k) - n\Big(\sum \ell = 1 J p_\ell - 1\Big). \tag{4.38}$$

This dual approach leads to the following proposition.

Proposition 4.1 *The EM algorithm for mixtures is equivalent to iteration (4.38).*

Proof. From (4.34) and (4.37), we have

$$p_\ell = \frac{1}{n}\sum_{i=1}^{n} t_{i\ell}(\theta^k) \tag{4.39}$$

for all $\ell = 1, \ldots, J$, which coincide with the values obtained with the EM algorithm. On the other hand, in view of (4.38), μ_ℓ and Σ_ℓ maximize $Q(\theta, \theta^k)$, exactly as in EM, independently of the constraint. □

We now turn to CEM^2.

Proposition 4.2 *The CEM² recursion is equivalent to a coordinate-wise generalized proximal point procedure of the type*

$$\theta^{k+1} = \arg\max_{\theta\in\Theta_k} \left\{ L(\theta \mid \boldsymbol{y}) - n(\sum \ell = 1Jp_\ell - 1) - D\big(\theta, \theta^k \mid y\big) \right\}, \tag{4.40}$$

where Θ_k is the parameter set of the form

$$\Theta_k = \left\{ \theta \in \mathbb{R}^p \mid \theta_\ell = \theta_\ell^k,\ \ell \neq j \right\}$$

with $j = k - \frac{k}{J}\rfloor J + 1$.

Proof. Looking at the maximization steps (3.22) and (2.8) and using formulation (4.38) for EM, we can easily deduce that, at iteration k of CEM², θ_j^{k+1} is equal to the jth component of

$$\arg\max_{\theta\in\mathbb{R}^p} \left\{ Q(\theta \mid \theta^k) - n(\sum \ell = 1Jp_\ell - 1) \right\} .$$

Then it is enough to note that $Q(\theta \mid \theta^k) - n(\sum \ell = 1Jp_\ell - 1)$ can be decomposed into

$$\sum \ell = 1J \left(Q_\ell(\theta_\ell \mid \theta^k) - n(p_\ell - \frac{1}{J}) \right) , \tag{4.41}$$

where $Q_\ell(\theta_\ell \mid \theta^k) = \sum i = 1nt_{i\ell}(\theta^k) \log p_\ell \varphi(y_i \mid \alpha_\ell)$. Each term of the sum in (4.41) only depends on θ_ℓ so that maximizing (4.41) in θ is equivalent to maximizing in the θ_ℓ's independently. Therefore θ_j^{k+1} is equal to the jth component of

$$\arg\max_{\theta\in\Theta_k} \left\{ Q(\theta \mid \theta^k) - n(\sum \ell = 1Jp_\ell - 1) \right\} .$$

Using (2.18), (4.40) is easily deduced for the jth component and the proof of the proposition is achieved since for $\ell \neq j$, CEM² clearly satisfies (4.40). □

Properties of the Kullback "semi-distance"

We begin with the following simple fact. Consider the quantity $D(\theta, \theta' \mid \mathbf{y})$ defined in (2.16). Since the Kullback-Leibler divergence is strictly convex, nonnegative and is zero between identical distributions, D vanishes iff $t(\theta') = t(\theta)$. However, the operator defined by $t(.)$ is not injective on the whole parameter space. Therefore, the Kullback information does not *a priori* behave like a distance in all directions of the parameter space. In the following lemma, we prove that $t(.)$ is coordinate-wise injective which allows the Kullback measure to enjoy some "distance like" properties at least on coordinate subspaces.

Lemma 4.3 *For any ν in $\{1, \ldots, J\}$ the operator $t(\theta_1, \ldots, \theta_{\nu-1}, ., \theta_{\nu+1}, \ldots, \theta_J)$ is injective.*

Proof. Fix ν in $\{1, \ldots, J\}$. Let $v = (p_\nu^v, \mu_\nu^v, \Sigma_\nu^v)$ and $w = (p_\nu^w, \mu_\nu^w, \Sigma_\nu^w)$ be two proposed vectors for the νth component such that

$$t(\theta_1, \ldots, \theta_{\nu-1}, v, \theta_{\nu+1}, \ldots, \theta_J) = t(\theta_1, \ldots, \theta_{\nu-1}, w, \theta_{\nu+1}, \ldots, \theta_J).$$

Define

$$\theta^v = (\theta_1, \ldots, \theta_{\nu-1}, v, \theta_{\nu+1}, \ldots, \theta_J)^T$$

and

$$\theta^w = (\theta_1, \ldots, \theta_{\nu-1}, w, \theta_{\nu+1}, \ldots, \theta_J)^T$$

and $(p_j^v, \mu_j^v, \Sigma_j^v)$ (resp. $(p_j^w, \mu_j^w, \Sigma_j^w)$), the components of θ^v (resp. θ^w) for $j = 1, \ldots, J$. Then, for any $j' \neq \nu$, for $i = 1, \ldots, n$,

$$t_{ij'}(\theta^v) = t_{ij'}(\theta^w) \; .$$

Since $j' \neq \nu$, the numerators of both terms in the last equation are equal. Thus, so are the denominators. Therefore

$$\sum_{j=1}^{J} p_j^v \varphi(y_i \mid \mu_j^v, \Sigma_j^v) = \sum_{j=1}^{J} p_j^w \varphi(y_i \mid \mu_j^w, \Sigma_j^w) \; .$$

Since all terms corresponding to $j \neq \nu$ are equal in the sums above, it follows straightforwardly that for $i = 1, \ldots, n$

$$p_\nu^v \varphi(y_i \mid \mu_\nu^v, \Sigma_\nu^v) = p_\nu^w \varphi(y_i \mid \mu_\nu^w, \Sigma_\nu^w).$$

For Gaussian mixtures, and for most mixtures of interest (exponential, binomial, Poisson, ...), these equations imply that v=w as soon as $n > 2$, ensuring that the operator $t(\theta_1, \ldots, \theta_{\nu-1}, ., \theta_{\nu+1}, \ldots, \theta_J)$ is injective. For Gaussian mixtures, for example, this comes from the fact that a polynom of order 2 is the null function as soon as it has more than two roots. □

¿From this lemma and the Kullback-Leibler divergence properties, the lemma below follows straightforwardly.

Lemma 4.4 *The distance-like function $D(\theta, \theta' \mid \boldsymbol{y})$ satisfies the following properties*
(i) $D(\theta, \theta' \mid \boldsymbol{y}) \geqslant 0$ for all θ' and θ in Θ,
(ii) if θ and θ' only differ in one coordinate, $D(\theta, \theta' \mid \boldsymbol{y}) = 0$ implies $\theta' = \theta$.

5 Convergence of CEM^2

Assumptions 5.1 *Let θ be any point in $\mathbb{R}^p$. Then, the level set*

$$\mathcal{L}_\theta = \left\{ \theta' \mid L(\theta' | \mathbf{y}) \geqslant L(\theta | \mathbf{y}) \right\} \tag{5.42}$$

is compact.

Let $\Lambda(\theta \mid \mathbf{y})$ be the modified log-likelihood function given by

$$\Lambda(\theta \mid \mathbf{y}) = L(\theta \mid \mathbf{y}) - n(\sum \ell = 1 J p_\ell - 1). \tag{5.43}$$

This function first arised in the Lagrangian framework of Section 4. It is indeed the Lagrangian function $\mathcal{L}(\theta, \lambda)$ taken at the value $\lambda = n$. We now establish a series of results concerning the CEM^2 iterations.

Proposition 5.1 *The sequence $\{\Lambda(\theta^k \mid \boldsymbol{y})\}_{k \in \mathbb{N}}$ is monotone non-decreasing, and satisfies*

$$\Lambda(\theta^{k+1} \mid \boldsymbol{y}) - \Lambda(\theta^k \mid \boldsymbol{y}) \geqslant D(\theta^{k+1}, \theta^k \mid \boldsymbol{y}). \tag{5.44}$$

Proof. ¿From iteration (4.40), we have

$$\Lambda(\theta^{k+1} \mid \mathbf{y}) - \Lambda(\theta^k \mid \mathbf{y}) \geqslant D(\theta^{k+1}, \theta^k \mid \mathbf{y}) - D(\theta^k, \theta^k \mid \mathbf{y}).$$

The proposition follows from $D(\theta^{k+1}, \theta^k \mid \mathbf{y}) \geqslant 0$ and $D(\theta^k, \theta^k \mid \mathbf{y}) = 0$. □

Lemma 5.2 *The sequence* $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ *is bounded and satisfies*

$$\lim_{k\to\infty}\sum_{j=1}^{J} p_j^k = 1 \tag{5.45}$$

If in addition, $\{\Lambda(\theta^k|\mathbf{y})\}_{k\in\mathbb{N}}$ *is bounded from above,*

$$\lim_{k\to\infty}\|\theta^{k+1}-\theta^k\| = 0\ . \tag{5.46}$$

Proof. The fact that $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ is bounded is straightforward from Proposition 5.1 and Assumption 5.1.

To show (i), we consider the sequence $\{\sum_{j=1}^{J} p_j^k\}_{k\in\mathbb{N}}$ and denote by $\{\sum_{j=1}^{J} p_j^{\sigma(k)}\}_{k\in\mathbb{N}}$ a converging subsequence. Let $\{\theta^{\sigma(\gamma(k))}\}_{k\in\mathbb{N}}$ be a converging subsequence of $\{\theta^{\sigma(k)}\}_{k\in\mathbb{N}}$ with θ^* its limit point. Using (3.22), it is easy to check that, for $j=1,\ldots,J$,

$$\lim_{k\to\infty} p_j^{\sigma(\gamma(k))} = \frac{1}{n}\sum_{i=1}^{n} t_{ij}(\theta^*)\ ,$$

from which it follows directly that

$$\lim_{k\to\infty}\sum_{j=1}^{J} p_j^{\sigma(\gamma(k))} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n} t_{ij}(\theta^*) = 1\ .$$

It comes that $\{\sum_{j=1}^{J} p_j^{\sigma(k)}\}_{k\in\mathbb{N}}$ converges necessarily to 1. Therefore (5.45) is satisfied for any such converging subsequence, which proves (i) since $\{\sum_{j=1}^{J} p_j^k\}_{k\in\mathbb{N}}$ is bounded.

The proof for 5.46 is similar. Considering a converging subsequence $\{\|\theta^{\sigma(k)+1}-\theta^{\sigma(k)}\|\}_{k\in\mathbb{N}}$ of the bounded sequence $\{\|\theta^{k+1}-\theta^k\|\}_{k\in\mathbb{N}}$, it is possible to extract a subsequence $\{\theta^{\sigma(\gamma(k))}\}_{k\in\mathbb{N}}$ such that $\{\theta^{\sigma(\gamma(k))}\}_{k\in\mathbb{N}}$ and $\{\theta^{\sigma(\gamma(k))+1}\}_{k\in\mathbb{N}}$ respectively converge to θ^{**} and θ^*. In addition, inequality (5.44) in Proposition 5.1 and convergence of $\left\{\Lambda(\theta^k \mid \mathbf{y})\right\}_{k\in\mathbb{N}}$ imply that

$$\lim_{k\to\infty} D(\theta^{k+1},\theta^k \mid \mathbf{y}) = 0. \tag{5.47}$$

By continuity of D, it comes $D(\theta^{**},\theta^* \mid \mathbf{y}) = 0$. Since θ^* and θ^{**} only differ in one coordinate, it follows from Lemma 4.4 that $\theta^* = \theta^{**}$. Then

$$\lim_{k\to\infty}\|\theta^{\sigma(\gamma(k))+1}-\theta^{\sigma(\gamma(k))}\| = \|\theta^{**}-\theta^*\| = 0\ ,$$

from which 5.46 follows easily. □

Theorem 5.3 *Every accumulation point* θ^* *of the sequence* $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ *satifies one of the following two properties*

- $\Lambda(\theta^* \mid \boldsymbol{y}) = +\infty$
- θ^* *is a stationary point of the modified log-likelihood function* $\Lambda(\theta \mid \boldsymbol{y})$.

Proof. Two cases are thus to be considered. In the first case, $\{\Lambda(\theta^k \mid \mathbf{y}))\}_{k\in\mathbb{N}}$ is unbounded. Since this sequence is increasing, $\lim_{k\to\infty} \Lambda(\theta^k \mid \mathbf{y}) = +\infty$. Moreover, since $\{\theta^k\}_{k\in\mathbb{N}}$ is bounded, one deduces that any accumulation point θ^* maximizes Λ with $\Lambda(\theta^* \mid \mathbf{y}) = +\infty$.

Let us now assume that $\{\Delta(\theta^k|\mathbf{y})\}_{k\in\mathbb{N}}$ is bounded from above. For any j in $\{1,\ldots,J\}$ consider the following subsequence $\left\{\theta^{\sigma(k)}\right\}_{k\in\mathbb{N}}$ of $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ such that

$$\sigma(k) - \frac{\sigma(k)}{J}\rfloor + 1 = j.$$

Since $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ is bounded, one can extract a converging subsequence $\left\{\theta^{\sigma(\gamma(k))}\right\}_{k\in\mathbb{N}}$ from $\left\{\theta^{\sigma(k)}\right\}_{k\in\mathbb{N}}$ with limit θ^*. The defining iteration (4.40) implies that

$$\frac{\partial}{\partial\theta_j}\Lambda(.\mid\mathbf{y})|_{\theta^{\sigma(\gamma(k))+1}} - \frac{\partial}{\partial\theta_j}D(.,\theta^{\sigma(\gamma(k))}\mid\mathbf{y})|_{\theta^{\sigma(\gamma(k))+1}} = 0.$$

Due to continuous differentiability of $\Lambda(.\mid\mathbf{y})$ and $D(.,.\mid\mathbf{y})$, the partial derivative of $\Lambda(\theta\mid\mathbf{y})$ is continuous in θ and the partial derivative of $D(\theta,\theta'\mid\mathbf{y})$ in the variable θ is continuous with respect to (θ,θ'). Hence, 5.46 in Lemma 5.2 gives

$$\frac{\partial}{\partial\theta_j}\Lambda(.\mid\mathbf{y})|_{\theta^*} - \frac{\partial}{\partial\theta_j}D(.,\theta^*\mid\mathbf{y})|_{\theta^*} = 0 \tag{5.48}$$

for all $j = 1,\ldots,J$. On the other hand, since $D(.,\theta^*\mid\mathbf{y})$ attains its minimum at θ^*, we have for all $j = 1,\ldots,J$

$$\frac{\partial}{\partial\theta_j}D(.,\theta^*\mid\mathbf{y})|_{\theta^*} = 0\,.$$

Thus, equation (5.48) gives, for all $j = 1,\ldots,J$

$$\frac{\partial}{\partial\theta_j}\Lambda(.\mid\mathbf{y})|_{\theta^*} = 0,$$

which concludes the proof. □

Corollary 5.4 *Assume that the constraint log-likelihood function $\Lambda(\theta\mid\boldsymbol{y})$ is strictly concave in an open neighborhood of a stationary point of $\left\{\theta^k\right\}_{k\in\mathbb{N}}$. Then, the sequence $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ converges and its limit is a local maximizer of $\Lambda(\theta\mid\boldsymbol{y})$.*

Proof. This is a direct consequence of Corollary 4.5 in [34]. □

We now prove the main convergence result for the CEM^2 procedure.

Theorem 5.5 *Every accumulation point of the sequence $\left\{\theta^k\right\}_{k\in\mathbb{N}}$ is a stationary point of the log-likelihood function $L(\theta\mid\mathbf{y})$ over the set defined by the constraint $\sum_{\ell=1}^{J} p_\ell = 1$.*

Proof. Let θ^* be an accumulation point of $\{\theta^k\}_{k\in\mathbb{N}}$. Consider the affine submanifold of Θ

$$\Theta' = \Big\{\theta\in\Theta \mid \sum_{\ell=1}^{J} p_\ell = 1\Big\}. \tag{5.49}$$

Notice that θ^* lies in Θ'. Take any vector δ such that $\theta^*+\delta$ lies in Θ'. Since Θ' is affine, any point $\theta_t = \theta^* + t\delta$, $t\in\mathbb{R}$ also lies in Θ'. The directional derivative of Λ at θ^* in the direction δ is obviously null. It is given by

$$(0 =)\Lambda'(\theta^*;\delta\mid\mathbf{y}) = \lim_{t\to 0^+}\frac{\Lambda(\theta^*\mid\mathbf{y}) - \Lambda(\theta^*+t\delta\mid\mathbf{y})}{t}, \tag{5.50}$$

which is equal to

$$\Lambda'(\theta^*;\delta \mid \mathbf{y}) = \lim_{t\to 0^+} \frac{L(\theta^* \mid \mathbf{y}) - L(\theta^* + t\delta \mid \mathbf{y}) + c(\theta^*) - c(\theta^* + t\delta)}{t}, \tag{5.51}$$

where $c(\theta) = n\Big(\sum_{\ell=1}^{J} p_\ell - 1\Big)$. Since $\theta^* + t\delta$ lies in Θ' for all nonnegative t, $c(\theta^* + t\delta) = c(\theta^*) = 0$, and we obtain

$$\Lambda'(\theta^*;\delta \mid \mathbf{y}) = L'(\theta^*;\delta \mid \mathbf{y}). \tag{5.52}$$

Thus,

$$L'(\theta^*;\delta \mid \mathbf{y}) = 0 \tag{5.53}$$

6 Numerical experiments

The behaviors of EM, SAGE (as described in the Appendix) and CEM^2 are compared on the basis of simulation experiments on univariate Gaussian mixtures with $J = 3$ components. First, we consider a mixture of well separated components with equal mixing proportions $p_1 = p_2 = p_3 = 1/3$, means $\mu_1 = 0, \mu_2 = 3, \mu_3 = 6$ and equal variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$. We will refer to this mixture as the well-separated mixture. Secondly, we consider a mixture of overlapping components with equal mixing proportions $p_1 = p_2 = p_3 = 1/3$, means $\mu_1 = 0, \mu_2 = 3, \mu_3 = 3$ and variances $\sigma_1^2 = \sigma_2^2 = 1, \sigma_3^2 = 4$. This mixture will be referred to as the overlapping mixture.

For the well-separated mixture we consider a unique sample of size $n = 300$ and perform the EM, SAGE and CEM^2 algorithms from the following initial position:

$$p_1^0 = p_2^0 = p_3^0 = 1/3,\ \ \mu_1^0 = \bar{x} - s, \mu_2^0 = \bar{x}, \mu_3^0 = \bar{x} + s, \sigma_1^0 = \sigma_2^0 = \sigma_3^0 = s^2$$

where $\bar{x}$ and s^2 are respectively the empirical sample mean and variance. Starting from this rather favorable initial position, close to the true parameter values, the three algorithms converge to the same solution below

$$\begin{aligned} \hat{p}_1 &= 0.36, \hat{\mu}_1 = 0.00, \hat{\sigma}_1^2 = 1.10, \\ \hat{p}_2 &= 0.29, \hat{\mu}_2 = 2.96, \hat{\sigma}_2^2 = 0.38 \\ \hat{p}_3 &= 0.35, \hat{\mu}_3 = 5.90, \hat{\sigma}_3^2 = 1.10 \end{aligned}$$

The performances of EM, SAGE and CEM^2, in terms of speed, are compared on the basis of the cycles number needed to reach the stationary value of the constraint log-likelihood. A cycle corresponds to the updating of all mixture components. For EM, it consists of a E-step (2.7) and a M-step (2.8). For SAGE, it is the (J+1) iterations described in the Appendix. For CEM^2, it consists of J iterations described in (3.21) and (3.22). In each case the algebraic operations needed to achieve a cycle of iterations are of the same nature and are in the same number, namely, J updatings of the mixing proportions, means and variance matrices and $J \times n$ updatings of the conditional probabilities $t_{ij}(\theta)$.

Figure B.1 displays the log-likelihood versus cycle for EM, SAGE and CEM^2 in the well-separated mixture case. As expected, when starting from a good initial position in a well separated mixture situation, EM converges rapidly to a local maximum of the likelihood. Moreover, EM outperforms SAGE and CEM^2 in this example.

For the overlapping mixture, we consider two different samples of size $n = 300$ and performed the EM, SAGE and CEM^2 algorithms from the following initial position:

$$p_1^0 = p_2^0 = p_3^0 = 1/3,\ \ \mu_1^0 = 0.0, \mu_2^0 = 0.1, \mu_3^0 = 0.2, \sigma_1^0 = \sigma_2^0 = \sigma_3^0 = 1.0,$$

Figure B.1: Comparison of log-likelihood versus cycle for EM (full line), SAGE (dashed line) and CEM^2 (dotted line) in the well-separated mixture case.

Figure B.2: Comparison of log-likelihood versus cycle for EM (full line), SAGE (dashed line) and CEM^2 (dotted line) in the overlapping mixture case (first sample).

Figure B.3: Comparison of log-likelihood versus cycle for EM (full line), SAGE (dashed line) and CEM^2 (dotted line) in the overlapping mixture case (second sample).

which is far from the true parameter values. For the first sample, the three algorithms converge to the same solution

$$\begin{aligned}\hat{p}_1 &= 0.65, \hat{\mu}_1 = 0.85, \hat{\sigma}_1^2 = 1.28\\ \hat{p}_2 &= 0.19, \hat{\mu}_2 = 3.32, \hat{\sigma}_2^2 = 0.26\\ \hat{p}_3 &= 0.16, \hat{\mu}_3 = 5.67, \hat{\sigma}_3^2 = 2.10\end{aligned}$$

Figure B.2 displays the log-likelihood versus cycle for EM, SAGE and CEM^2 for the first sample of the overlapping mixture. In this situation, EM appears to converge slowly so that SAGE and especially CEM^2 show a significant improvement of convergence speed.

For the second sample, starting from the same position, SAGE and CEM^2 both converge to the solution below

$$\begin{aligned}\hat{p}_1 &= 0.61, \hat{\mu}_1 = 0.85, \hat{\sigma}_1^2 = 1.62\\ \hat{p}_2 &= 0.13, \hat{\mu}_2 = 3.00, \hat{\sigma}_2^2 = 0.52\\ \hat{p}_3 &= 0.26, \hat{\mu}_3 = 4.27, \hat{\sigma}_3^2 = 4.29,\end{aligned}$$

while EM proposes the following solution, after 1000 cycles,

$$\begin{aligned}\hat{p}_1 &= 0.61, \hat{\mu}_1 = 0.83, \hat{\sigma}_1^2 = 1.60\\ \hat{p}_2 &= 0.16, \hat{\mu}_2 = 2.98, \hat{\sigma}_2^2 = 0.62\\ \hat{p}_3 &= 0.22, \hat{\mu}_3 = 4.58, \hat{\sigma}_3^2 = 4.29.\end{aligned}$$

Figure B.3 displays the log-likelihood versus cycle for EM, SAGE and CEM^2 for the second sample of the overlapping mixture. The same conclusions hold for this sample. The CEM^2 algorithm is the faster while EM is really slow, the correspondant local maximum of the likelihood not being reached after 1000 ierations.

Moreover, it appears that the implemented version of the SAGE algorithm is less beneficial than CEM^2 for situations in which EM converges slowly. A possible reason for this median behavior of SAGE is that the $(J+1)$th iteration of SAGE involves the whole complete data structure, whereas CEM^2 iterations never need the whole complete data space as hidden data space.

7 Discussion

We presented a component-wise EM algorithm for finite identifiable mixtures of distributions (CEM^2) and proved convergence properties similar to that of standard EM. As illustrated in section 6, numerical experiments suggest that CEM^2 and EM have complementary performances. The CEM^2 algorithm is of poor interest when EM convergence is fast but shows significant improvement when EM encounters slow convergence rate. Thus, CEM^2 may be most useful in many contexts. An intuitive explanation of our procedure performances is that the component-wise strategy prevents the algorithm from staying too long at critical points (typically saddle points) where standard EM is likely to get trapped. More theoretical investigations would be interesting but are beyond the scope of the present paper.

Other futur directions of research include the use of relaxation, as in [33], for accelerating CEM^2, and the possibility of using varying/adaptative orders to update the components.

Appendix: The SAGE algorithm

The Space-Alternating Generalized EM (SAGE) algorithm [17] is one of the most promising general purpose extention of EM. The SAGE method aims to avoid slow convergence situations of the EM method by updating small groups of the elements of the parameter vector associated to small hidden data spaces rather than one large complete data space. General description and details concerning the rationale, the properties and illustrations of the SAGE algorithm can be found in [17], [16], [15].

In this section, we restrict to maximum likelihood estimation for incomplete data parametric models.

We consider an incomplete data model where the parameter vector θ lies in a subset Θ of $\mathbb{R}^p$. (For a general multivariate Gaussian mixture, we have $p = (J-1)+Jd+Jd(d+1)/2$.) Let S be a non empty subset of $\{1,\ldots,p\}$ and $\tilde{S}$ its complement. We denote by θ_S the parameter components with indices in S. In order to describe the SAGE algorithm, we need the following definition.

Definition: a random vector X^s with probability density function $\mathbf{f}(\mathbf{x}^s \mid \theta)$ is an *admissible hidden-data space* with respect to θ_s for $\mathbf{g}(\mathbf{y} \mid \theta)$ if the joint density of X^s and Y satisfies

$$\mathbf{f}(\mathbf{y}, \mathbf{x}^s \mid \theta) = \mathbf{f}(\mathbf{y} \mid \mathbf{x}^s, \theta_{\tilde{s}})\mathbf{f}(\mathbf{x}^s \mid \theta), \tag{7.54}$$

i.e. the conditional distribution $\mathbf{f}(\mathbf{y} \mid \mathbf{x}^s, \theta)$ must be independent of θ_s.

Let $\theta^0 \in \Theta$ be an initial parameter estimate. The kth iteration of the SAGE algorithm is as follows.

(a) Choose an index set S^k.

(b) Choose an admissible hidden-data space X^{S^k}

(c) **E-step**: Compute

$$Q^{S^k}(\theta_{S^k} \mid \theta^k) = I\!E\left[\log \mathbf{f}(\mathbf{x}^{s^k} \mid \theta_{S^k}, \theta^k_{\tilde{S}^k}) \mid \mathbf{y}, \theta^k\right].$$

(d) **M-step**: Find

$$\begin{aligned} \theta^{k+1}_{S^k} &= \arg\max_{\theta_{S^k}} Q^{S^k}(\theta_{S^k} \mid \theta^k), \\ \theta^{k+1}_{\tilde{S}^k} &= \theta^k_{\tilde{S}^k}. \end{aligned} \tag{7.55}$$

Fessler and Hero [17] showed convergence properties of the SAGE algorithm analogous to that of the EM algorithm. Moreover, they showed that the asymptotic rate of the SAGE algorithm is improved if one chooses a less informative hidden data space. Numerous numerical experiments [17], [16] support this assertion.

As noted in [17], choosing index sets is as much art as science. In practical situations, a SAGE algorithm can be decomposed in cycles of iterations according to repeated choices of the index sets S. In the Gaussian mixture context, we propose choosing the index sets as described below.

A SAGE algorithm cycle is composed of $(J+1)$ iterations. In the first J iterations, $j = 1,\ldots,J$, the chosen index set S^j contains the indices of the mean vector μ_j and variance

matrix Σ_j of the jth mixture component. The associated hidden-data space is $Y \times Z^j$, where $Z^j = (Z_i^j, i = 1, \ldots, n)$, $Z_i^\ell \in \{0, 1\}$ being the random variable indicating whether y_i arises from the jth component.

The E-step consists of computing

$$Q(\mu_j, \Sigma_j | \theta^{k+\frac{j-1}{J+1}}) = \sum_{i=1}^{n} t_{ij}(\theta^{k+\frac{j-1}{J+1}}) \log \varphi(y_i | \mu_j, \Sigma_j),$$

and reduces to update the conditional probabilities, given y_i, that unit i arises from component j for $i = 1, \ldots, n$. For $j = 1, \ldots, J$, we introduce the notation $\alpha_j = (\mu_j, \Sigma_j)$, and consider that $k - 1 + j/(J + 1) = 0$ if $k = 0$. Therefore, the jth iteration ($j = 1, \ldots, J$) consists of the following E and M steps.

Compute

$$t_{ij}^{k+\frac{j}{J+1}}(\theta^{k+\frac{j-1}{J+1}}) = \frac{p_j^{k+\frac{j-1}{J+1}} \varphi(y_i \mid \alpha_j^{k+\frac{j-1}{J+1}})}{\sum_{\ell<j} p_\ell^{k+\frac{\ell}{J+1}} \varphi(y_i \mid \alpha_\ell^{k+\frac{\ell}{J+1}}) + \sum_{\ell \geq j} p_\ell^{k-1+\frac{\ell}{J+1}} \varphi(y_i \mid \alpha_\ell^{k-1+\frac{\ell}{J+1}})}.$$

and set for $\ell \neq j$ and $i = 1, \ldots, n$

$$t_{i\ell}^{k+\frac{j}{J+1}}(\theta^{k+\frac{j-1}{J+1}}) = t_{i\ell}^{k+\frac{\ell}{J+1}}(\theta^{k+\frac{\ell-1}{J+1}}) \text{ if } \ell < j,$$

$$t_{i\ell}^{k+\frac{j}{J+1}}(\theta^{k+\frac{j-1}{J+1}}) = t_{i\ell}^{k-1+\frac{\ell}{J+1}}(\theta^{k-1+\frac{\ell-1}{J+1}}) \text{ if } \ell > j.$$

The M-step consists of maximizing in μ_j and Σ_j the function $Q(\mu_j, \Sigma_j | \theta^{k+\frac{j-1}{J+1}})$ and leads to

$$\begin{aligned}
\mu_j^{k+\frac{j}{J+1}} &= \frac{\sum_{i=1}^{n} t_{ij}^{k+\frac{j}{J+1}}(\theta^{k+\frac{j-1}{J+1}}) y_i}{\sum_{i=1}^{n} t_{ij}^{k+\frac{j}{J+1}}(\theta^{k+\frac{j-1}{J+1}})} \\
\Sigma_j^{k+\frac{j}{J+1}} &= \frac{\sum_{i=1}^{n} t_{ij}^{k+\frac{j}{J+1}}(\theta^{k+\frac{j-1}{J+1}})(y_i - \mu_j^{k+\frac{j}{J+1}})(y_i - \mu_j^{k+\frac{j}{J+1}})'}{\sum_{i=1}^{n} t_{ij}^{k+\frac{j}{J+1}}(\theta^{k+\frac{j-1}{J+1}})}.
\end{aligned}$$

All the other parameter estimates are unchanged. Namely, for $\ell = 1, \ldots, J$,

$$p_\ell^{k+\frac{j}{J+1}} = p_\ell^{k+\frac{j-1}{J+1}}$$

and for $\ell \neq j$

$$\begin{aligned}
\mu_\ell^{k+\frac{j}{J+1}} &= \mu_\ell^{k+\frac{j-1}{J+1}} \\
\Sigma_\ell^{k+\frac{j}{J+1}} &= \Sigma_\ell^{k+\frac{j-1}{J+1}}.
\end{aligned}$$

The $(J + 1)$th iteration concerns the mixing proportions. The index set is the indices of the mixing proportions $p_1, \ldots, p_J$. The associated hidden-data space is the whole complete data space $Y \times Z$, where $(Z = Z^1, \ldots, Z^J)$.

The E-step of that iteration consits of computing

$$Q(p_1,\ldots,p_J|\theta^{k+\frac{J}{J+1}}) = \sum_{i=1}^{n}\sum_{j=1}^{J} t_{ij}(\theta^{k+\frac{J}{J+1}})\log p_j,$$

which , for $j = 1,\ldots,J$ and $i = 1,\ldots,n$, reduces to the computation of

$$t_{ij}^{k+1}(\theta^{k+\frac{J}{J+1}}) = \frac{p_j^{k+\frac{J}{J+1}}\varphi(y_i \mid \alpha_j^{k+\frac{J}{J+1}})}{\sum_{\ell=1}^{J} p_\ell^{k+\frac{J}{J+1}}\varphi(y_i \mid \alpha_\ell^{k+\frac{J}{J+1}})}.$$

The M-step of the $(J+1)$th iteration consists of updating the mixing proportions, for $j = 1,\ldots,J$,

$$p_j^{k+1} = \frac{\sum_{i=1}^{n} t_{ij}^{k+1}(\theta^{k+\frac{J}{J+1}})}{n},$$

letting the other parameter estimates unchanged.

As already mentioned in Section 3, this choice of the SAGE algorithm is not fully component-wise since the mixing proportions are updated in the same iteration. The reason why it is not possible to deal with the mixing proportions separately is that the maximization of (7.55) cannot be ensured since the constraint (2.9) cannot be fulfilled. Notice that our CEM^2 algorithm has been conceived in the same spirit as the present SAGE algorithm, but it is not a SAGE algorithm since the updating steps of mixing proportions cannot be regarded as maximisation steps of the form (7.55).

Bibliography

[1] Ciarlet, Philippe G., Introduction to numerical linear algebra and optimization. With the assistance of Bernadette Miara and Jean-Marie Thomas for the exercises. Transl. by A. Buttigieg., Cambridge Texts in Applied Mathematics : Cambridge University Press., (1988). (p. 68).

[2] Chretien, S. and Hero, A. O., Acceleration of the EM algorithm via proximal point iterations, IEEE International Symposium on Information Theory, MIT Boston, (1998). (p. 63).

[3] Dempster, A. P. and Laird, N. M. and Rubin, D. B., Maximum likelihood for incomplete data via the EM algorithm (with discussion), J. Roy. Stat. Soc. Ser. B, 39 (1977), 1–38, (p. 63).

[4] Redner, R. A. and Walker, H. F., Mixture densities, maximum lilelihood and the EM algorithm, SIAM Review, 26, (1984) 195–239. (pp. 63, 64 et 70).

[5] Martinet, B., Régularisation d'inéquation variationnelles par approximations successives, Revue Francaise d'Informatique et de Recherche Operationnelle, 3, (1970), 154–179. (p. 44).

[6] Ostrowski, A. M., Solution of equations and systems of equations, Academic, New York, 1966

[7] Bonnans, J. F. and Gilbert, J.-Ch. and Lemaréchal, C. and Sagastizàbal, C., Optimization numérique. Aspects théoriques et pratiques, Series : Mathématiques et Applications, 27, Springer Verlag, 1997. (pp. 53 et 55).

[8] Hiriart Urruty, J. B. and Lemaréchal, C., Convex Analysis and Minimization Algorithms I-II, Grundlehren der mathematischen Wissenschaften 306, Springer Verlag, 1993.

[9] Rockafellar, R. T., Monotone operators and the proximal point algorithm, SIAM. J. CONT. OPT., 14, (1976), 877–898. (pp. 44, 47, 63 et 67).

[10] Rockafellar, R. T., Augmented Lagrangians and application of the proximal point algorithm in convex programming, Mathematics of Operations Research, 1, (1976), 96–116.

[11] Teboulle, M., Entropic proximal mappings with application to nonlinear programming, Mathematics of Operations Research, 17, (1992) 670–690. (pp. 44, 67 et 96).

[12] Teboulle, M., Convergence of proximal-like algorithms, SIAM Journal On Optimization, 7, (1997) 1069–1083. (pp. 64 et 67).

[13] Bauschke, H. H. and Borwein, J. M., On Projection Algorithms for Solving Convex Feasibility Problems, SIAM REVIEW, 38 (1996) 3, 367–426. (p. 50).

[14] Wu, C. F., "On the convergence of the EM algorithm", Ann. Statis., 11, (1983), 1, 95–103

[15] Hero, A. O. and Fessler, J. A., Convergence in norm for EM-type algorithms, Statistica Sinica, 5, (1995), 1, 41–54. (pp. 63 et 80).

[16] Fessler, J. A. and Hero, A. O., Penalized maximum-likelihood image reconstruction using space-Alternating generalized EM algorithms, IEEE Trans. Image Processing, 4, (1995), 1417–1429.

[17] Fessler, J. A. and Hero, A. O., Space-Alternating generalized expectation-maximisation algorithm, IEEE Trans. Signal Processing, 42, (1994), 2664–2677. (pp. 63, 64 et 80).

[18] Lewitt, R. M. and Muehllehner, Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum likelihood estimation, IEEE Tr. Med. Im., 5, (1986), 1, 16–22. (pp. 63, 64, 69 et 80).

[19] Kaufman, L., Implementing and accelerating the EM algorithm for positron emission tomography, IEEE Tr. Med. Im., 6 (1987), 1, 37–51. (p. 63).

[20] Jamshidian, M. and Jennrich, R. I., Conjugate gradient acceleration of the EM algorithm, J. Am. Stat. Ass., 88, (1993), 421, 221–228. (p. 63).

[21] Meilijson, I., A fast improvement to the EM algorithm in its own terms, J. Roy. Stat. Soc. Ser. B, 51, (1989), 127–138. (p. 63).

[22] Louis, T. A., Finding the observed information matrix when using the EM algorithm, J. Roy. Stat. Soc. Ser. B, 44, (1982), 226–233. (p. 63).

[23] Meng, X.-L., T. A. and van Dyk, D. A., The EM algorithm - an old folk song sung to a fast new tune (with discussion), J. Roy. Stat. Soc. Ser. B, 59, (1997), 511–567. (p. 63).

[24] Meng, X.-L., T. A. and van Dyk, D. A., Fast EM-type implementations for mixed effects models, J. Roy. Stat. Soc. Ser. B, 60, (1998), 559–578. (p. 63).

[25] Hathaway, R.J., Another interpretation of EM algorithm for mixture distributions, Statist. and Probab. Letters, 4 (1986), 53–56. (p. 63).

[26] Bouman, C. and Sauer, K., A unified approach to statistical tomography using coordinate descent optimization, Proc. 27th Conf. Info. Sci. Sys., John Hopkins, (1993), 611–616 (p. 66).

[27] Lange, K. and Carson, R., EM reconstruction algorithms for emission and transmission tomography, J. Comp. Assisted Tomo., 8 (1984), 2, 306–316.

[28] Lange, K. and Carson, R., EM reconstruction algorithms for emission and transmission tomography, J. Comp. Assisted Tomo., 8 (1984)2, 306–316.

[29] Moré, J. J., Recent developments in algorithms and software for trust region methods, Mathematical Programming: The State of the Art, Bonn, Springer Verlag, 258–287, 1983.

[30] Ibragimov, I. A. and Khas'minskij, R. Z., Asymptotic theory of estimation, Teoriya Veroyatnostej i Matematicheskaya Statistika. Moskva: Nauka, Glavnaya Redaktsiya Fiziko-Matematicheskoj Literatury, 1979. (p. 67).

[31] Luenberger, D. G., Optimization by vector space methods (Series in Decision and Control), New York-London. Sydney-Toronto: John Wiley and Sons, Inc., 1969.

[32] McLachlan, G. J. and Krishnam, T., The EM algorithm and extensions,New York-London. Sydney-Toronto: John Wiley and Sons, Inc., 1997. (p. 63).

[33] Chrétien, S and Hero, A. O., Kullback proximal algorithms for maximum likelihood estimation, Technical Report, CSPL, The University of Michigan, Ann Arbor, USA,1998. (pp. 63, 67 et 79).

[34] Chrétien, S and Hero, A. O., Generalized proximal point algorithms and bundle implementation, Technical Report, CSPL 313, The University of Michigan, Ann Arbor, USA, 1998. (pp. 63, 67 et 75).

[35] Neal, R. N. and Hinton, G. E., A view of the EM algorithm that justifies incremental, sparse and other variants, Learning in Graphical Models, Jordan, M.I. (Editor), Dordrencht, Kluwer Academic Publishers, (1998), 195–239 (pp. 63 et 66).

[36] Lange, K., A quasi-newtonian acceleration of the EM algorithm, Statistica Sinica, 5, (1995) 670-690.

[37] Lange, K., A gradient algorithm locally equivalent to the EM algorithm, J. Roy. Stat. Soc. Ser. B, 52, (1995), 425-337,

[38] Titterington, D. M., Smith, A. F. M. and Makov, U. E., Statistical analysis of finite mixture distributions, Wiley, New York, 1985.

[39] Auslender, A., Optimisation. Méthodes numériques, Masson, Paris, 1976.

[40] Liu, C. and Sun, D. X., Acceleration of EM algorithm for mixtures models using ECME, ASA Proceedings of The Stat. Comp. Session, (1997), 109-114. (p. 63).

[41] Meng, X. L. and Rubin, D. B., Maximum likelihood estimation via the ECM algorithm: A general framework, Biometrika, 80, (1993). (p. 63).

[42] Liu, C. and Rubin, D.B., The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence, Biometrika, 81, (1994) 633-648. (p. 63).

[43] Liu, C. and Rubin, D.B. and Wu, Y., Parameter expansion to accelerate EM: the PX-EM algorithm, Biometrika, (1998), 755-770. (p. 63).

[44] Pilla, R. S. and Lindsay, B. G., Alternative EM methods in high-dimensional finite mixtures, Technical report, Department of Statistics, Penn State University (1996). (p. 63).

Chapter C

On EM algorithms and their proximal generalizations

with Alfred O. Hero.

Abstract

In this paper, we analyze the celebrated EM algorithm from the point of view of proximal point algorithms. More precisely, we study a new type of generalization of the EM procedure introduced in [4] and called Kullback-proximal algorithms. The proximal framework allows us to prove new results concerning the cluster points. An essential contribution is a detailed analysis of the case where some cluster points lie on the boundary of the parameter space.

1 Introduction

The problem of maximum likelihood (ML) estimation consists of finding a solution of the form

$$\theta_{ML} = argmax_{\theta\in\Theta}\, l_y(\theta), \tag{1.1}$$

where y is an observed sample of a random variable Y defined on a sample space $\mathcal{Y}$ and $l_y(\theta)$ is the log-likelihood function defined by

$$l_y(\theta) = \log g(y;\theta), \tag{1.2}$$

defined on the parameter space $\Theta \subset \mathbb{R}^n$, and $g(y;\theta)$ denotes the density of Y at y parametrized by the vector parameter θ.

The Expectation Maximization (EM) algorithm is an iterative procedure which is widely used for solving ML estimation problems. The EM algorithm was first proposed by Dempster, Laird and Rubin [8] and has seen the number of its potential applications increase substantially since its appearance. The book of McLachlan and Krishnan [14] gives a comprehensive overview of the theoretical properties of the method and its applicability.

The convergence of the sequence of EM iterates towards a maximizer of the likelihood function was claimed in the original paper [8] but it was later noticed that the proof contained a flaw. A careful convergence analysis was finally given by Wu [21] based on Zangwill's general theory [23]; see also [14]. Zangwill's theory applies to general iterative schemes and the main task when using it is to verify that the assumptions of Zangwill's theorems

are satisfied. Since the appearance of Wu's paper, convergence of the EM algorithm is often taken for granted in many cases where the necessary assumptions were sometimes not carefully justified. As an example, an often neglected issue is the behavior of EM iterates when they approach the boundary of the domain of definition of the functions involved. A different example is the following. It is natural to try and establish that EM iterates actually converge to a single point θ^*, which involves proving uniqueness of the cluster point. Wu's approach, reported in [14, Theorem 3.4, p. 89] is based on the assumption that the euclidean distance between two successive iterates tends to zero. However such an assumption is in fact very hard to verify in most cases and should not be deduced solely from experimental observations.

The goal of the present paper is to propose an analysis of EM iterates and their generalizations in the framework of Kullback proximal point algorithms. We focus on the geometric conditions that are provable in practice and the concrete difficulties concerning convergence towards boundaries and cluster point uniqueness. The approach adopted here was first proposed in [4] in which it was shown that the EM algorithm could be recast as a Proximal Point algorithm. A proximal scheme for maximizing the function $l_y(\theta)$ using the distance-like function I_y is an iterative procedure of the form

$$\theta^{k+1} \in argmax_{\theta\in\Omega} l_y(\theta) - \beta_k I_y(\theta, \theta^k), \tag{1.3}$$

where $(\beta_k)_{k\in\mathbb{N}}$ is a sequence of positive real numbers often called relaxation parameters. Proximal point methods were introduced by Martinet [13] and Rockafellar [17] in the context of convex minimization. The proximal point representation of the EM algorithm [4] is obtained by setting $\beta_k = 1$ and $I_y(\theta, \theta^k)$ to the Kullback distance between some well specified conditional densities of a complete data vector. The general case of $\beta_k > 0$ was called the Kullback Proximal Point algorithm (KPP). This approach was further developed in [5] where convergence was studied in the twice differentiable case with the assumption that the limit point lies in the interior of the domain. The main novelty of [5] was to prove that relaxation of the Kullback-type penalty could ensure superlinear convergence which was confirmed by experiment for a Poisson linear inverse problem. This paper is an extension of these previous works that addresses the problem of convergence under general conditions.

The main results of this paper are the following. Firstly, we prove that all the cluster points of the Kullback proximal sequence which lie in the interior of the domain are stationary points of the likelihood function l_y under very mild assumptions that are easily verified in practice. Secondly, taking into account finer properties of I_y, we prove that every cluster point on the boundary of the domain satisfies the Karush-Kuhn-Tucker necessary conditions for optimality under nonnegativity constraints. To illustrate our results, we apply the Kullback-proximal algorithm to an estimation problem in animal carcinogenicity introduced in [1] in which an interesting nonconvex constraint is handled. In this case, the M-step cannot be obtained in closed form. However, the Kullback-proximal algorithm can be analyzed and implemented. Numerical experiments are provided which demonstrate the ability of the method to significantly accelerate the convergence of standard EM.

The paper is organized as follows. In Section 2, we review the Kullback proximal point interpretation of EM. Then, in Section 3 we study the properties of interior cluster points. We prove that such cluster points are in fact global maximizers of a certain penalized likelihood function. This allows us to justify using a relaxation parameter β when β is sufficiently small to permit avoiding saddle points. Section 4 pursues the analysis in the case where the cluster point lies on a boundary of the domain of I_y.

2 The Kullback proximal framework

In this section, we review the EM algorithm and the Kullback proximal interpretation discussed in [5].

The EM algorithm

The EM procedure is an iterative method which produces a sequence $(\theta^k)_{k\in\mathbb{N}}$ such that each θ^{k+1} maximizes a local approximation of the likelihood function in the neighborhood of θ^k. This point of view will become clear in the proximal point framework of the next subsection.

In the traditional approach, one assumes that some data are hidden from the observer. A frequent example of hidden data is the class to which each sample belongs in the case of mixtures estimation. Another example is when the observed data are projection of an unkown object as for image reconstruction problems in tomography. One would prefer to consider the likelihood of the complete data instead of the ordinary likelihood. Since some parts of the data are hidden, the so called complete likelihood cannot be computed and therefore must be approximated. For this purpose, we will need some appropriate notations and assumptions which we now describe. The observed data are assumed to be i.i.d. samples from a unique random vector Y taking values on a data space $\mathcal{Y}$. Imagine that we have at our disposal more informative data than just samples from Y. Suppose that the more informative data are samples from a random variable X taking values on a space $\mathcal{X}$ with density $f(x;\theta)$ also parametrized by θ. We will say that the data X is more informative than the actual data Y in the sense that Y is a compression of X, i.e. there exists a non-invertible transformation h such that $Y = h(X)$. If one had access to the data X it would therefore be advantageous to replace the ML estimation problem (3.1) by

$$\hat{\theta}_{ML} = argmax_{\theta\in\mathbb{R}^p} l_x(\theta), \tag{2.4}$$

with $l_x(\theta) = \log f(x;\theta)$. Since $y = h(x)$ the density g of Y is related to the density f of X through

$$g(y;\theta) = \int_{h^{-1}(\{y\})} f(x;\theta)d\mu(x) \tag{2.5}$$

for an appropriate measure μ on $\mathcal{X}$. In this setting, the data y are called *incomplete data* whereas the data x are called *complete data*.

Of course the complete data x corresponding to a given observed sample y are unknown. Therefore, the complete data likelihood function $l_x(\theta)$ can only be estimated. Given the observed data y and a previous estimate of θ denoted $\bar{\theta}$, the following minimum mean square error estimator (MMSE) of the quantity $l_x(\theta)$ is natural

$$Q(\theta,\bar{\theta}) = E[\log f(x;\theta)|y;\bar{\theta}],$$

where, for any integrable function $F(x)$ on $\mathcal{X}$, we have defined the conditional expectation

$$E[F(x)|y;\bar{\theta}] = \int_{h^{-1}(\{y\})} F(x)k(x|y;\bar{\theta})d\mu(x)$$

and $k(x|y;\bar{\theta})$ is the conditional density function given y

$$k(x|y;\bar{\theta}) = \frac{f(x;\bar{\theta})}{g(y;\bar{\theta})}. \tag{2.6}$$

Having described the notions of complete data and complete likelihood and its local estimation we now turn to the EM algorithm. The idea is relatively simple: a legitimate way to proceed is to require that iterate θ^{k+1} be a maximizer of the local estimator of the complete likelihood conditionally on y and θ^k. Hence, the EM algorithm generates a sequence of approximations to the solution (2.4) starting from an initial guess θ^0 of θ_{ML} and is defined by

$$\textbf{Compute } Q(\theta, \theta^k) = E[\log f(x;\theta)|y;\theta^k] \qquad \textbf{E Step}$$

$$\theta^{k+1} = argmax_{\theta\in\mathbb{R}^p} Q(\theta, \theta^k) \qquad \textbf{M Step}$$

Kullback proximal interpretation of the EM algorithm

Consider the general problem of maximizing a concave function $\Phi(\theta)$. The original proximal point algorithm introduced by Martinet [13] is an iterative procedure which can be written

$$\theta^{k+1} = argmax_{\theta\in D_\Phi} \left\{ \Phi(\theta) - \frac{\beta_k}{2}\|\theta - \theta^k\|^2 \right\}. \tag{2.7}$$

The quadratic penalty $\frac{1}{2}\|\theta - \theta^k\|^2$ is relaxed using a sequence of positive parameters $\{\beta_k\}$. In [17], Rockafellar showed that superlinear convergence of this method is obtained when the sequence $\{\beta_k\}$ converges towards zero.

It was proved in [5] that the EM algorithm is a particular example in the class of proximal point algorithms using Kullback Leibler types of penalties. One proceeds as follows. Assume that the family of conditional densities $\{k(x|y;\theta)\}_{\theta\in\mathbb{R}^p}$ is regular in the sense of Ibragimov and Khasminskii [9], in particular $k(x|y;\theta)\mu(x)$ and $k(x|y;\bar\theta)\mu(x)$ are mutually absolutely continuous for any θ and $\bar\theta$ in $\mathbb{R}^p$. Then the Radon-Nikodym derivative $\frac{k(x|y,\bar\theta)}{k(x|y;\theta)}$ exists for all $\theta, \bar\theta$ and we can define the following Kullback Leibler divergence:

$$I_y(\theta, \bar\theta) = E\big[\log \frac{k(x|y,\bar\theta)}{k(x|y;\theta)}|y;\bar\theta\,\big]. \tag{2.8}$$

We are now able to define the Kullback-proximal algorithm. For this purpose, let us define D_l as the domain of l_y, $D_{I,\theta}$ the domain of $I_y(\cdot,\theta)$ and D_I the domain of $I_y(\cdot,\cdot)$.

Definition 2.1 *Let $(\beta_k)_{k\in\mathbb{N}}$ be a sequence of positive real numbers. Then, the Kullback-proximal algorithm is defined by*

$$\theta^{k+1} = argmax_{\theta\in D_l\cap D_{I,\theta^k}} l_y(\theta) - \beta_k I_y(\theta, \theta^k). \tag{2.9}$$

The main result on which the present paper relies is that EM algorithm is a special case of (2.6), i.e. it is a penalized ML estimator with proximal penalty $I_y(\theta, \theta^k)$.

Proposition 2.2 *[5, Proposition 1] The EM algorithm is a special instance of the Kullback-proximal algorithm with $\beta_k = 1$, for all $k \in \mathbb{N}$.*

The previous definition of the Kullback proximal algorithm may appear overly general to the reader familiar with the usual practical interpretation of the EM algorithm. However, we found that such a framework has at least the three following benefits [5]:

- to our opinion, the convergence proof of our EM is more natural,

- the Kullback proximal framework may easily incorporate additional constraints, a feature that may be of crucial importance as demonstrated in the example of Section 5 below,
- the relaxation sequence $(\beta_k)_{k\in\mathbb{N}}$ allows one to weight the penalization term and its convergence to zero implies quadratic convergence in certain examples.

The first of these three arguments is also supported by our simplified treatment of the componentwise EM procedure proposed in [3] and the remarkable recent results of [20] based on a special proximal entropic representation of EM for getting precise estimates on the convergence speed of EM algorithms, however, with much more restrictive assumptions than the ones of the present paper.

Although our results are obtained under mild assumptions concerning the relaxation sequence $(\beta_k)_{k\in\mathbb{N}}$ including the case $\beta_k = 0$, several precautions should be taken when implementing the method. However, one of the key features of EM-like procedures is to allow easy handling of positivity or more complex constraints, such as the ones discussed in the example of Section 5. In such cases the function I_y behaves like a barrier whose value increases to infinity as the iterates approach the boundary of the constraint set. Hence, the sequence $(\beta_k)_{k\in\mathbb{N}}$ ought to be positive in order to exploit this important computational feature. On the other hand, as proved under twice differentiability assumptions in [5] when the cluster set reduces to a unique nondegenerate maximizer in the interior of the domain of the log-likelihood and β_k converges to zero, quadratic convergence is obtained. This nice behavior is not satisfied in the plain EM case where $\beta_k = 1$ for all $k \in \mathbb{N}$. As a drawback, one problem in decreasing the β_k's too quickly is possible numerical ill conditioning. The problem of choosing the relaxation sequence is still largely open. We have found however that for most "reasonable" sequences, our method was at least as fast as the standard EM.

Finally, we would like to end our presentation of KPP-EM by noting that closed form iterations may not be available in the case $\beta_k \neq 1$. If this is the case, solving (2.6) becomes a subproblem which will require iterative algorithms. In some interesting examples, e.g. the case presented in Section 5. In this case, the standard EM iterations are not available in closed form in the first place and KPP-EM provides faster convergence while preserving monotonicity and constraint satisfaction.

Notations and assumptions

The notation $\|\cdot\|$ will be used to denote the norm on any previously defined space without more precision. The space on which it is the norm should be obvious from the context. For any bivariate function Φ, $\nabla_1\Phi$ will denote the gradient with respect to the first variable. In the remainder of this paper we will make the following assumptions.

Assumptions 2.1 *(i) l_y is differentiable on D_l and $l_y(\theta)$ tends to $-\infty$ whenever $\|\theta\|$ tends to $+\infty$.*
(ii) the projection of D_I onto the first coordinate is a subset of D_l.
(iii) $(\beta_k)_{k\in\mathbb{N}}$ is a convergent nonnegative sequence of real numbers whose limit is denoted by β^*.

We will also impose the following assumptions on the distance-like function I_y.

Assumptions 2.2 *(i) There exists a finite dimensional euclidean space S, a differentiable mapping $t : D_l \mapsto S$ and a functional $\Psi : D_\Psi \subset S \times S \mapsto \mathbb{R}$ such that*

$$I_y(\theta, \bar{\theta}) = \Psi(t(\theta), t(\bar{\theta})),$$

where D_ψ denotes the domain of Ψ.
(ii) For any $\{t^k, t)_{k\in\mathbb{N}}\} \subset D_\Psi$ there exists $\rho_t > 0$ such that $\lim_{\|t^k - t\|\to\infty} I_y(t^k, t) \geqslant \rho_t$. Moreover, we assume that $\inf_{t\in M} \rho_t > 0$ for any bounded set $M \subset S$.
For all (t', t) in D_Ψ, we will also require that
(iii) (Positivity) $\Psi(t', t) \geqslant 0$,
(iv) (Identifiability) $\Psi(t', t) = 0 \Leftrightarrow t = t'$,
(v) (Continuity) Ψ is continuous at (t', t)
and for all t belonging to the projection of D_Ψ onto its second coordinate,
(vi) (Differentiability) the function $\Psi(\cdot, t)$ is differentiable at t.

Assumptions 2.2(i) and (ii) on l_y are standard and are easily checked in practical examples, e.g. they are satisfied for the Poisson and additive mixture models. Notice that the domain D_I is now implicitly defined by the knowledge of D_l and D_Ψ. Moreover I_y is continuous on D_I. The importance of requiring that I_y has the prescribed shape comes from the fact that I_y might not satisfy assumption 2.2(iv) in general. Therefore assumption 2.2 (iv) reflects the requirement that I_y should at least satisfy the identifiability property up to a possibly injective transformation. In both examples discussed above, this property is an easy consequence of the well known fact that $a\log(a/b) = 0$ implies $a = b$ for positive real numbers a and b. The growth, continuity and differentiability properties 2.2 (ii), (v) and (vi) are, in any case, nonrestrictive.

For the sake of notational convenience, the regularized objective function with relaxation parameter β will be denoted

$$F_\beta(\theta, \bar\theta) = l_y(\theta) - \beta I_y(\theta, \bar\theta). \tag{2.10}$$

Finally we make the following general assumption.

Assumptions 2.3 *The Kullback proximal iteration (2.6) is well defined, i.e. there exists at least one maximizer of $F_{\beta^k}(\theta, \theta^k)$ at each iteration k.*

In the EM case, i.e. $\beta = 1$, this last assumption is equivalent to the computability of M-steps. A sufficient condition for this assumption to hold would be, for instance, that $F_\beta(\theta, \bar\theta)$ be sup-compact, i.e. the level sets $\{\theta \mid F_\beta(\theta, \bar\theta) \geqslant \alpha\}$ be compact for all α, $\beta > 0$ and $\bar\theta \in D_l$. However, this assumption is not usually satisfied since the distance-like function is not defined on the boundary of its domain. In practice it suffices to solve the equation $\nabla F_{\beta^k}(\theta, \theta^k) = 0$, to prove that the solution is unique. Then assumption 2.2(i) is sufficient to conclude that we actually have a maximizer.

General properties : monotonicity and boundedness

Using Assumptions 2.2, we easily deduce monotonicity of the likelihood values and boundedness of the proximal sequence. The first two lemmas are proved, for instance, in [5].

We start with the following monotonicity result.

Lemma 2.3 *[5, Proposition 2] For any iteration $k \in \mathbb{N}$, the sequence $(\theta^k)_{k\in\mathbb{N}}$ satisfies*

$$l_y(\theta^{k+1}) - l_y(\theta^k) \geqslant \beta_k I_y(\theta^k, \theta^{k+1}) \geqslant 0. \tag{2.11}$$

From the previous lemma, we easily obtain the boundedness of the sequence.

Lemma 2.4 *[5, Lemma 2] The sequence $(\theta^k)_{k\in\mathbb{N}}$ is bounded.*

The next lemma will also be useful.

Lemma 2.5 *Assume that there exists a subsequence $(\theta^{\sigma(k)})_{k\in\mathbb{N}}$ belonging to a compact set C included in D_l. Then,*

$$\lim_{k\to\infty} \beta_k I_y(\theta^{k+1}, \theta^k) = 0.$$

Proof. Since l_y is continuous over C, $\sup_{\theta\in C} l_y(\theta) < +\infty$ and $(l_y(\theta^{\sigma(k)}))_{k\in\mathbb{N}}$ is therefore bounded from above. Moreover, Lemma 3.1 implies that the sequence $(l_y(\theta^k))_{k\in\mathbb{N}}$ is monotone nondecreasing. Therefore, the whole sequence $(l_y(\theta^k))_{k\in\mathbb{N}}$ is bounded from above and convergent. This implies that $\lim_{k\to\infty} l_y(\theta^{k+1}) - l_y(\theta^k) = 0$. Applying Lemma 3.1 again, we obtain the desired result. □

3 Analysis of interior cluster points

The convergence analysis of Kullback proximal algorithms is split into two parts, the first part being the subject of this section. We prove that if the accumulation points θ^* of the Kullback proximal sequence satisfy $(\theta^*, \theta^*) \in D_{I_y}$ they are stationary points of the log-likelihood function l_y. It is also straightforward to show that the same analysis applies to the case of penalized likelihood estimation.

Nondegeneracy of the Kullback penalization

We start with the following useful lemma.

Lemma 3.1 *Let $(\alpha_1^k)_{k\in\mathbb{N}}$ and $(\alpha_2^k)_{k\in\mathbb{N}}$ be two bounded sequences in D_Ψ satisfying*

$$\lim_{k\to\infty} \Psi(\alpha_1^k, \alpha_2^k) = 0.$$

Assume that every couple (α_1^*, α_2^*) of accumulation points of these two sequences lies in D_Ψ. Then,

$$\lim_{k\to\infty} \|\alpha_1^k - \alpha_2^k\| = 0.$$

Proof. First, one easily obtains that $(\alpha_2^k)_{k\in\mathbb{N}}$ is bounded (use a contradiction argument and Assumption 2.2 (ii)). Assume that there exits a subsequence $(\alpha_1^{\sigma(k)})_{k\in\mathbb{N}}$ such that $\|\alpha_1^{\sigma(k)} - \alpha_2^{\sigma(k)}\| \geqslant 3\varepsilon$ for some $\varepsilon > 0$ and for all large k. Since $(\alpha_1^{\sigma(k)})_{k\in\mathbb{N}}$ is bounded, one can extract a convergent subsequence. Thus we may assume without any loss of generality that $(\alpha_1^{\sigma(k)})_{k\in\mathbb{N}}$ is convergent with limit α^*. Using the triangle inequality, we have $\|\alpha_1^{\sigma(k)} - \alpha_1^*\| + \|\alpha_1^* - \alpha_2^{\sigma(k)}\| \geqslant 3\varepsilon$. Since $(\alpha_1^{\sigma(k)})_{k\in\mathbb{N}}$ converges to α_1^*, there exists a integer K such that $k \geqslant K$ implies $\|\alpha_1^{\sigma(k)} - \alpha_1^*\| \leqslant \varepsilon$. Thus for $k \geqslant K$ we have $\|\alpha_1^* - \alpha_2^{\sigma(k)}\| \geqslant 2\varepsilon$. Now recall that $(\alpha_2^k)_{k\in\mathbb{N}}$ is bounded and extract a convergent subsequence $(\alpha_2^{\sigma(\gamma(k))})_{k\geqslant K}$ with limit denoted by α_2^*. Then, using the same arguments as above, we obtain $\|\alpha_1^* - \alpha_2^*\| \geqslant \varepsilon$. Finally, recall that $\lim_{k\to\infty} \Psi(\alpha_1^k, \alpha_2^k) = 0$. We thus have $\lim_{k\to\infty} \Psi(\alpha_1^{\sigma(\gamma(k))}, \alpha_2^{\sigma(\gamma(k))}) = 0$, and, due to the fact that the sequences are bounded and $\Psi(\cdot,\cdot)$ is continuous in both variables, we have $I_y(\alpha_1^*, \alpha_2^*) = 0$. Thus assumption 2.2 (iv) implies that $\|\alpha_1^* - \alpha_2^*\| = 0$ and we obtain a contradiction. Hence, $\lim_{k\to\infty} \|\alpha_1^k - \alpha_2^k\| = 0$ as claimed. □

Cluster points

The main results of this section are the following. First, we prove that under the assumptions 2.2, 2.2 and 2.7, any cluster point θ^* is a global maximizer of $F_{\beta^*}(\theta^*, \theta^*)$. We then use this general result to prove that such cluster points are stationary points of the log-likelihood

function. This result motivates a natural assumption under which θ^* is in fact a local maximizer of l_y. In addition we show that if the sequence $(\beta^k)_{k\in\mathbb{N}}$ converges to zero, i.e. $\beta^* = 0$, then θ^* is a global maximizer of log-likelihood. Finally, we discuss some simple conditions under which the algorithm converges, i.e. has only one cluster point.

The following theorem states a result which describes the stationary points of the proximal point algorithm as global maximizers of the asymptotic penalized function.

Theorem 3.2 *Assume that $\beta^* > 0$. Let θ^* be any accumulation point of $(\theta^k)_{k\in\mathbb{N}}$. Assume that $(\theta^*, \theta^*) \in D_I$. Then, θ^* is a global maximizer of the penalized function $F_{\beta^*}(\cdot, \theta^*)$ over the projection of D_I onto its first coordinate, i.e.*

$$F_{\beta^*}(\theta^*, \theta^*) \geqslant F(\theta, \theta^*)$$

for all θ such that $(\theta, \theta^*) \in D_I$.

An informal argument is as follows. Assume that $\Theta = \mathbb{R}^n$. From the definition of the proximal iterations, we have

$$F_{\beta_{\sigma(k)}}(\theta^{\sigma(k)+1}, \theta^{\sigma(k)}) \geqslant F_{\beta_{\sigma(k)}}(\theta, \theta^{\sigma(k)})$$

for all subsequence $(\theta^{\sigma(k)})_{k\in\mathbb{N}}$ converging to θ^* and for all $\theta \in \Theta$. Now, assume we can prove that $\theta^{\sigma(k)}$ also converges to θ^*, we obtain by taking the limit and using continuity, that

$$F_{\beta_*}(\theta^*, \theta^*) \geqslant F_{\beta_*}(\theta, \theta^*)$$

which is the required result. There are two major difficulties when one tries to transform this sketch into a rigorous argument. The first one is related to the fact that l_y and I_y are only defined on domains which may not to be closed. Secondly, proving that $\theta^{\sigma(k)}$ converges to θ^* is not an easy task. This issue will be discussed in more detail in the next section. The following proof overcomes both difficulties.

Proof. Without loss of generality, we may reduce the analysis to the case where $\beta_k \geqslant \beta > 0$ for a certain β. The fact that θ^* is a cluster point implies that there is a subsequence of $(\theta^k)_{k\in\mathbb{N}}$ converging to θ^*. For k sufficiently large, we may assume that the terms $(\theta^{\sigma(k+1)}, \theta^{\sigma(k)})$ belong to a compact neighborhood C^* of (θ^*, θ^*) included in D_I. Recall that

$$F_{\beta_{\sigma(k)-1}}(\theta^{\sigma(k)}, \theta^{\sigma(k)-1}) \geqslant F_{\beta_{\sigma(k)-1}}(\theta, \theta_{\sigma(k)-1})$$

for all θ such that $(\theta, \theta^{\sigma(k)-1}) \in D_I$ and a fortiori for $(\theta, \theta^{\sigma(k)-1}) \in C^*$. Therefore,

$$\begin{aligned} F_{\beta^*}(\theta^{\sigma(k)}, \theta^{\sigma(k)-1}) \quad & -(\beta_k - \beta^*) I_y(\theta^{\sigma(k)}, \theta^{\sigma(k)-1}) \geqslant \\ & F_{\beta^*}(\theta, \theta^{\sigma(k)-1}) - (\beta_{\sigma(k)-1} - \beta^*) I_y(\theta, \theta^{\sigma(k)-1}). \end{aligned} \tag{3.12}$$

Let us have a precise look at the "long term" behavior of I_y. First, since $\beta_k > \beta_*$ for all k sufficiently large, Lemma 3.3 says that

$$\lim_{k\to\infty} I_y(\theta^{\sigma(k)}, \theta^{\sigma(k)-1}) = 0.$$

Thus, for any $\varepsilon > 0$, there exits an integer K_1 such that $I_y(\theta^{\sigma(k)}, \theta^{\sigma(k)+1}) \leqslant \varepsilon$ for all $k \geqslant K_1$. Moreover, Lemma 3.1 and continuity of t allows to conclude that

$$\lim_{k\to\infty} t(\theta^{\sigma(k)-1}) = t(\theta^*).$$

Since Ψ is continuous, for all $\varepsilon > 0$ and for all k sufficienlty large we have

$$\begin{aligned} I_y(\theta^*, \theta^*) &= \Psi(t(\theta^*), t(\theta^*)) \\ &\geqslant \Psi(t(\theta^{\sigma(k)}), t(\theta^{\sigma(k)-1})) - \varepsilon \\ &= I_y(\theta^{\sigma(k)}, \theta^{\sigma(k)-1}) - \varepsilon. \end{aligned} \tag{3.13}$$

On the other hand, F_{β^*} is continuous in both variables on C^*, due to Assumptions 2.2(i) and 2.2(i). By continuity in the first and second arguments of $F_{\beta^*}(\cdot, \cdot)$, for any $\varepsilon > 0$ there exists $K_2 \in \mathbb{N}$ such that for all $k \geqslant K_2$

$$F_{\beta^*}(\theta^*, \theta) \leqslant F_{\beta^*}(\theta^{\sigma(k)}, \theta) + \varepsilon. \tag{3.14}$$

Using (3.13), since l_y is continuous, we obtain existence of K_3 such that for all $k \geqslant K_3$

$$F_{\beta^*}(\theta^*, \theta^*) \geqslant F_{\beta^*}(\theta^{\sigma(k)}, \theta^{\sigma(k)+1}) - 2\varepsilon. \tag{3.15}$$

Combining equations (3.14) and (3.15) with (3.9), we obtain

$$\begin{aligned} F_{\beta^*}(\theta^*, \theta^*) \geqslant \quad & F_{\beta^*}(\theta^*, \theta) - (\beta_k - \beta^*) I_y(\theta^{\sigma(k)}, \theta) \\ & + (\beta_k - \beta^*) I_y(\theta^{\sigma(k)}, \theta^{\sigma(k)+1})) - 3\varepsilon. \end{aligned} \tag{3.16}$$

Now, since $\beta^* = \lim_{k \to \infty} \beta_k$, there exists an integer K_4 such that $\beta_k - \beta^* \leqslant \varepsilon$ for all $k \geqslant K_4$. Therefore for all $k \geqslant \max\{K_1, K_2, K_3, K_4\}$, we obtain

$$F_{\beta^*}(\theta^*, \theta^*) \geqslant F_{\beta^*}(\theta^*, \theta) - \varepsilon I_y(\theta^{\sigma(k)}, \theta) - \varepsilon^2 - 3\varepsilon.$$

Since I_y is continuous and $(\theta^{\sigma(k)})_{k \in \mathbb{N}}$ is bounded, there exists a real constant K such that $I_y(\theta^{\sigma(k)}, \theta) \leqslant K$, for all $n \in \mathbb{N}$. Thus, for all k sufficiently large

$$F_{\beta^*}(\theta^*, \theta^*) \geqslant F_{\beta^*}(\theta^*, \theta) - (4\varepsilon K + \varepsilon^2). \tag{3.17}$$

Finally, recall that no assumption was made on θ, and that C^* is any compact neighborhood of θ^*. Thus, using the assumption 2.2(i), which asserts that $l_y(\theta)$ tends to $-\infty$ as $\|\theta\|$ tends to $+\infty$, we may deduce that (3.17) holds for any θ such that $(\theta, \theta^*) \in D_I$ and, letting ε tend to zero, we see that θ^* maximizes $F_{\beta^*}(\theta, \theta^*)$ for over all θ such that (θ, θ^*) belongs to D_I as claimed. □

Using this theorem, we may now deduce that certain accumulation points on the strict interior of the parameter's space are stationary points of the log-likelihood function.

Corollary 3.3 *Assume that $\beta^* > 0$. Let θ^* be any accumulation point of $(\theta^k)_{k \in \mathbb{N}}$. Assume that $(\theta^*, \theta^*) \in int D_I$. Then, if l_y is differentiable on D_l, θ^* is a stationary point of $l_y(\theta)$. Moreover, if l_y is concave, then θ^* is a global maximizer of l_y.*

Proof. Since under the required assumptions l_y is differentiable and $I_y(\theta^*, \cdot)$ is differentiable at θ^*, Theorem 3.2 states that

$$0 \in \Big\{ \nabla l_y(\theta^*) + \beta_* \nabla I_y(\theta^*, \theta^*) \Big\}.$$

Since $I_y(\cdot, \theta^*)$ is minimum at θ^*, $\nabla_1 I_y(\theta^*, \theta^*) = 0$ and we thus obtain that θ^* is a stationary point of l_y. This implies that θ^* is a global maximizer in the case where l_y is concave. □.

Theorem 3.2 seems to be much stronger than the previous corollary. The fact that accumulation points of the proximal sequence may not be global maximizers of the likelihood is now easily seen to be a consequence of fact that the Kullback distance-like function I_y perturbs the shape of the likelihood function when θ is far from θ^*. This perturbation does not have serious consequence in the concave case. On the other hand, one may wonder whether θ^* cannot be proved to be at least a local maximizer instead of a mere stationary point. The answer is given in the following corollary.

Corollary 3.4 *Let θ^* be an accumulation point of $(\theta^k)_{k\in\mathbb{N}}$ such that $(\theta^*,\theta^*) \in intD_I$. In addition, assume that l_y and $I_y(\cdot,\theta^*)$ are twice differentiable in a neighborhood of θ^* and that the Hessian matrix $\nabla^2 l_y(\theta^*)$ at θ^* is not the null matrix. Then, if β^* is sufficiently small, θ^* is a local maximizer of l_y over D_l.*

Proof. Assume that θ^* is not a local maximizer. Since $\nabla^2 l_y$ is not the null matrix, for β^* sufficiently small, there is a direction δ in the tangent space to D_l for which the function $f(t) = F_{\beta^*}(\theta^* + t\delta, \theta^*)$ has positive second derivative for t sufficiently small. This contradicts the fact that θ^* is a global maximizer of $F_{\beta^*}(\cdot,\theta^*)$ and the proof is completed. □

The next theorem establishes global optimality of accumulation points in the case where the relaxation sequence converges to zero.

Theorem 3.5 *Let θ^* be any accumulation point of $(\theta^k)_{k\in\mathbb{N}}$. Assume that $(\theta^*,\theta^*) \in D_I$. Then, without assuming differentiability of either l_y or of I_y, if $(\beta_k)_{k\in\mathbb{N}}$ converges to zero, θ^* is a global maximizer of l_y over the projection of D_I along the first coordinate.*

Proof. Let $(\theta^{\sigma(k)})_{k\in\mathbb{N}}$ be a convergent subsequence of $(\theta^k)_{k\in\mathbb{N}}$ with limit denoted θ^*. We may assume that for k sufficiently large, $(\theta^{\sigma(k+1)}, \theta^{\sigma(k)})$ belongs to a compact neighborhood C^* of θ^*. By continuity of l_y, for any $\varepsilon > 0$, there exists $K \in \mathbb{N}$ such that for all $k \geqslant K$,

$$l_y(\theta^*) \geqslant l_y(\theta^{\sigma(k)}) - \varepsilon.$$

On the other hand, the proximal iteration (1.3) implies that

$$l_y(\theta^{\sigma(k)}) - \beta_{\sigma(k)-1} I_y(\theta^{\sigma(k)-1}, \theta^{\sigma(k)}) \geqslant l_y(\theta) - \beta_{\sigma(k)-1} I_y(\theta^{\sigma(k)-1}, \theta),$$

for all $\theta \in D_l$. Fix $\theta \in D_l$. Thus, for all $k \geqslant K$,

$$l_y(\theta^*) \geqslant l_y(\theta) + \beta_{\sigma(k)-1} I_y(\theta^{\sigma(k)-1}, \theta^{\sigma(k)}) - \beta_{\sigma(k)-1} I_y(\theta^{\sigma(k)-1}, \theta) - \varepsilon.$$

Since I_y is a nonnegative function and $(\beta_k)_{k\in\mathbb{N}}$ is a nonnegative sequence, we obtain

$$l_y(\theta^*) \geqslant l_y(\theta) - \beta_{\sigma(k)-1} I_y(\theta^{\sigma(k)-1}, \theta) - \varepsilon.$$

Recall that $(\theta^k)_{k\in\mathbb{N}}$ is bounded due to Lemma 3.2. Thus, since I_y is continuous, there exists a constant C such that $I_y(\theta^{\sigma(k)-1}, \theta) \leqslant C$ for all k. Therefore, for k greater than K,

$$l_y(\theta^*) \geqslant l_y(\theta) - \beta_{\sigma(k)-1} C - \varepsilon.$$

Passing to the limit, and recalling that $(\beta_k)_{k\in\mathbb{N}}$ tends to zero, we obtain that

$$l_y(\theta^*) \leqslant l_y(\theta) - \varepsilon.$$

Using the same argument as at the end of the proof of Theorem 3.2, this latter equation holds for any θ such that (θ,θ^*) belongs to D_I, which concludes the proof upon letting ε tend to zero. □

Convergence of the Kullback proximal sequence

One question remains open in the analysis of the previous section: does the sequence generated by the Kullback proximal point converge? In other words: are there multiple cluster points? In Wu's paper [21], the answer takes the following form. If the euclidean distance between two successive iterates tends to zero, a well known result states that the set of accumulation points is a continuum (see for instance [16, Theorem 28.1]) and therefore, it is connected. Therefore, if the set of stationary points of l_y is a countable set, the iterates must converge.

Theorem 3.6 *Let S^* denote the set of accumulation points of the sequence $(\theta^k)_{k\in\mathbb{N}}$. Assume that $\lim_{k\to\infty} \|\theta^{k+1}-\theta^k\| = 0$ and that $l_y(\theta)$ is strictly concave in an open neighborhood $\mathcal{N}$ of an accumulation point θ^* of $(\theta^k)_{k\in\mathbb{N}}$ and that (θ^*,θ^*) is in $intD_I$. Then, for any relaxation sequence $(\beta_k)_{k\in\mathbb{N}}$, the sequence $(\theta^k)_{k\in\mathbb{N}}$ converges to a local maximizer of $l_y(\theta)$.*

Proof. We obtained in Corollary 3.3 that every accumulation point θ^* of $(\theta^k)_{k\in\mathbb{N}}$ in $intD_{l_y}$ and such that $(\theta^*,\theta^*) \in intD_{I_y}$ is a stationary point of $l_y(\theta)$. Since $l_y(\theta)$ is strictly concave over $\mathcal{N}$, the set of stationary points of l_y belonging to $\mathcal{N}$ reduces to singleton. Thus θ^* is the unique stationary point in $\mathcal{N}$ of l_y, and *a fortiori*, the unique accumulation point of $(\theta^k)_{k\in\mathbb{N}}$ belonging to $\mathcal{N}$. To complete the proof, it remains to show that there is no accumulation point in the exterior of $\mathcal{N}$. For that purpose, consider an open ball $\mathcal{B}$ of center θ^* and radius ε included in $\mathcal{N}$. Then, x^* is the unique accumulation point in $\mathcal{B}$. Moreover, any accumulation point θ', lying in the exterior of $\mathcal{N}$ must satisfy $\|\theta^*-\theta'\| \geqslant \varepsilon$, and we obtain a contradiction with the fact that S^* is connected. Thus every accumulation point lies in $\mathcal{N}$, from which we conclude that θ^* is the only accumulation point of $(\theta^k)_{k\in\mathbb{N}}$ or, in other words, that $(\theta^k)_{k\in\mathbb{N}}$ converges towards θ^*. Finally, notice that the strict concavity of $l_y(\theta)$ over $\mathcal{N}$ implies that θ^* is a local maximizer. □

Before concluding this section, let us make two general remarks.

- Proving *a priori* that the set of stationary points of l_y is discrete may be a hard task in specific examples.

- In general, it is not known whether $\lim_{k\to\infty} \|\theta^{k+1}-\theta^k\| = 0$ holds. In fact, Lemma 3.1 could be a first step in this direction. Indeed if we could prove in any application that the mapping t is injective, the desired result would follow immediately. However, injectivity of t does not hold in many of the standard examples; in the case of Gaussian mixtures, see [3, Section 2.2] for instance. Thus we are now able to clearly understand why the assumption that $\lim_{k\to\infty} \|\theta^{k+1}-\theta^k\| = 0$ is not easily deduced from general arguments. This problem has been overcome in [3] where it is shown that t is componentwise injective and thus performing a componentwise EM algorithm is a good alternative to the standard EM.

4 Analysis of cluster points on the boundary

The goal of this section is to extend the previous results to the case where some cluster points lie on the boundary of the region where computation of proximal steps is well defined. Such cluster points have rarely been analyzed in the statistical literature and the strategy developed for the interior case cannot be applied without further study of the Kullback distance-like function. Notice further that entropic-type penalization terms in proximal algorithms have been the subject of an intensive research effort in the mathematical programming community with the goal of handling positivity constraints; see [19] and the references therein for instance. The analysis proposed here applies to the more general Kullback distance-like functions I_y that occur in EM. Our goal is to show that such cluster points satisfy the well known Karush-Kuhn-Tucker conditions of nonlinear programming which extend the stationarity condition $\nabla l_y(\theta) = 0$ to the case where θ is subject to constraints. As before, it is straightforward to extend the proposed analysis to the case of penalized likelihood estimation.

In the sequel, the distance-like function will be assumed to have the following additional properties.

Assumptions 4.1 *The Kullback distance-like function I_y is of the form*

$$I_y(\theta,\bar{\theta}) = \sum_{1\leqslant i\leqslant n, 1\leqslant j\leqslant m} \alpha_{ij}(y_j) t_{ij}(\theta)\phi\Big(\frac{t_{ij}(\bar{\theta})}{t_{ij}(\theta)}\Big),$$

where for all i and j, t_{ij} is continuously differentiable on its domain of definition, α_{ij} is a function from $\mathcal{Y}$ to $\mathbb{R}_+$, the set of positive real numbers, and the function ϕ is a non negative convex continuously differentiable function defined for positive real numbers only and such that $\phi(\tau) = 0$ if and only if $\tau = 1$.

If $t_{ij}(\theta) = \theta_i$ and $\alpha_{ij} = 1$ for all i and all j, the function I_y is the well known ϕ divergence defined by Csiszàr in [7]. Assumption 2.4 is satisfied in most standard examples (for instance Gaussian mixtures and Poisson inverse problems) with the choice $\phi(\tau) = \tau\log(\tau)$.

More properties of the Kullback distance-like function

The main property that will be needed in the sequel is that under Assumption 2.4, the function I_y satisfies the same property as the one given in Lemma 3.1 above, even on the boundary of its domain D_I. This is the result of Proposition 3.4 below. We begin with one elementary lemma.

Lemma 4.1 *Under Assumptions 2.4, the function ϕ is decreasing on $(0,1)$, is increasing on $(1,+\infty)$ and $\phi(\tau)$ converges to $+\infty$ when τ converges to $+\infty$. We have $\lim_{k\to+\infty}\phi(\tau^k) = 0$ if and only if $\lim_{k\to+\infty}\tau^k = 1$.*

Proof. The first statement is obvious. For the second statement, the "if" part is trivial, so we only prove the "only if" part. First notice that the sequence $(\tau^k)_{k\in\mathbb{N}}$ must be bounded. Indeed, the level set $\{\tau \mid \phi(\tau) \leqslant \gamma\}$ is bounded for all $\gamma \geqslant 0$ and contains the sequence $(\tau^k)_{k\geqslant K}$ for K sufficiently large. Thus, the Bolzano-Weierstass theorem applies. Let τ^* be an accumulation point of $(\tau^k)_{k\in\mathbb{N}}$. Since ϕ is continuous, we get that $\phi(\tau^*) = 0$ and thus we obtain $\tau^* = 1$. From this, we deduce that the sequence has only one cluster point, which is equal to 1. Therefore, $\lim_{k\to+\infty}\tau^k = 1$. □

Using these lemmas, we are now in position to state and prove the main property of I_y.

Proposition 4.2 *The following statements hold.*

(i) For any sequence $(\theta^k)_{k\in\mathbb{N}}$ in $\mathbb{R}_+$ and any bounded sequence $(\eta^k)_{k\in\mathbb{N}}$ in $\mathbb{R}_+$, the fact that $\lim_{k\to+\infty} I_y(\eta^k,\theta^k) = 0$ implies $\lim_{k\to+\infty}|t_{ij}(\eta^k) - t_{ij}(\theta^k)| = 0$ for all i,j such that $\alpha_{ij} \neq 0$.

(ii) If one coordinate of one of the two sequences $(\theta^k)_{k\in\mathbb{N}}$ and $(\eta^k)_{k\in\mathbb{N}}$ tends to infinity, so does the other's same coordinate.

Proof. Fix i in $\{1,\dots,n\}$ and j in $\{1,\dots,m\}$ and assume that $\alpha_{ij} \neq 0$.

(i) We first assume that $(t_{ij}(\eta_i^k))_{k\in\mathbb{N}}$ is bounded away from zero.

Since $\lim_{k\to+\infty} I_y(\theta^k,\eta^k) = 0$, then $\lim_{k\to+\infty}\phi(t_{ij}(\theta^k)/t_{ij}(\eta^k)) = 0$ and Lemma 4.1 implies that $\lim_{k\to+\infty} t_{ij}(\theta^k)/t_{ij}(\eta^k) = 1$. Thus, $\lim_{k\to+\infty}(t_{ij}(\theta^k) - t_{ij}(\eta^k))/t_{ij}(\eta^k) = 0$ and since t is continuous, $t_{ij}(\eta^k)$ is bounded. This implies that $\lim_{k\to+\infty}|t_{ij}(\theta^k) - t_{ij}(\eta^k)| = 0$.

Next, consider the case of a subsequence $(t_{ij}(\eta^{\sigma(k)}))_{k\in\mathbb{N}}$ which tends towards zero. For contradiction, assume the existence of a subsequence $(t_{ij}(\theta^{\sigma(\gamma(k))})_{k\in\mathbb{N}}$ which remains bounded away from zero, i.e. there exists $a > 0$ such that $t_{ij}(\theta^{\sigma(\gamma(k))})_{k\in\mathbb{N}} \geqslant a$ for k sufficiently large. Thus, for k sufficiently large we get

$$\frac{t_{ij}(\theta^{\sigma(\gamma(k))})}{t_{ij}(\eta^{\sigma(\gamma(k))})} \geqslant \frac{a}{t_{ij}(\eta^{\sigma(\gamma(k))})} > 1,$$

and due to the fact that ϕ is increasing on $(1, +\infty)$, we obtain

$$t_{ij}(\eta^{\sigma(\gamma(k))})\phi\Big(\frac{t_{ij}(\theta^{\sigma(\gamma(k))})}{t_{ij}(\eta^{\sigma(\gamma(k))})}\Big) \geqslant t_{ij}(\eta^{\sigma(\gamma(k))})\phi\Big(\frac{a}{t_{ij}(\eta^{\sigma(\gamma(k))})}\Big). \tag{4.18}$$

On the other hand, Lemma 4.1 says that for any $b > 1$, $\phi'(b) > 0$. Since ϕ is convex, we get

$$\phi(\tau) \geqslant \phi(b) + \phi'(b)(\tau - b).$$

Take $\tau = a/t_{ij}(\eta^k)$ in this last expression and combine with (2.-6) to obtain

$$t_{ij}(\eta^{\sigma(\gamma(k))})\phi\Big(\frac{t_{ij}(\theta^{\sigma(\gamma(k))})}{t_{ij}(\eta^{\sigma(\gamma(k))})}\Big) \geqslant t_{ij}(\eta^{\sigma(\gamma(k))})(\phi(b) + \phi'(b)\Big(\frac{a}{t_{ij}(\eta^{\sigma(\gamma(k))})} - b\Big).$$

Passing to the limit, we obtain

$$0 = \lim_{k\to+\infty} t_{ij}(\eta^{\sigma(\gamma(k))})\phi\Big(\frac{t_{ij}(\theta^{\sigma(\gamma(k))})}{t_{ij}(\eta^{\sigma(\gamma(k))})}\Big) \geqslant a\phi'(b) > 0,$$

which gives the required contradiction.

(ii) If $(t_{ij}(\theta^k))_{k\in\mathbb{N}} \to +\infty$ then $(t_{ij}(\eta^k))_{k\in\mathbb{N}} \to +\infty$ is a direct consequence of part (i). Indeed, if $t_{ij}(\eta^k)$ remains bounded, part (i) says that $\lim_{k\to+\infty} |t_{ij}(\eta^k) - t_{ij}(\theta^k)| = 0$, which contradicts divergence of $(t_{ij}(\theta^k))_{k\in\mathbb{N}}$.

Now, consider the case where $(t_{ij}(\eta^k))_{k\in\mathbb{N}} \to +\infty$. Then, a contradiction is easily obtained if we assume that at least a subsequence $(t_{ij}(\theta^{\sigma(k)})_{k\in\mathbb{N}}$ stays bounded from above. Indeed, in such a case, we have

$$\lim_{k\to+\infty} \frac{t_{ij}(\theta^{\sigma(k)})}{t_{ij}(\eta^{\sigma(k)})} = 0,$$

and thus, $\phi(t_{ij}(\theta^k)/t_{ij}(\eta^k)) \geqslant \gamma$ for some $\gamma > 0$ since we know that ϕ is decreasing on $(0, 1)$ and $\phi(1) = 0$. This implies that

$$\lim_{k\to+\infty} t_{ij}(\eta^{\sigma(k)})\phi\Big(\frac{t_{ij}(\theta^{\sigma(k)})}{t_{ij}(\eta^{\sigma(k)})}\Big) = +\infty,$$

which is the required contradiction. □

Cluster points are KKT points

The main result of this section is the property that any cluster point θ^* such that (θ^*, θ^*) lies on the boundary of D_I satisfies the Karush-Kuhn-Tucker necessary conditions for optimality on the domain of the log-likelihood function. In the context of Assumptions 2.4, D_I is the set

$$D_I = \{\theta \in \mathbb{R}^n \mid t_{ij}(\theta) > 0 \quad \forall i \in \{1, \dots, n\} \text{ and } j \in \{1, \dots, m\}\}.$$

We have the following theorem.

Theorem 4.3 *Let θ^* be a cluster point of the Kullback-proximal sequence. Assume that all the functions t_{ij} are differentiable at θ^*. Let $\mathcal{I}^*$ be the set of all couples of indices (i, j) such that the constraint $t_{ij}(\theta) \geqslant 0$ is active at θ^*, i.e. $t_{ij}(\theta^*) = 0$. If θ^* lies in the interior of D_l, then θ^* satisfies the Karush-Kuhn-Tucker necessary conditions for optimality, i.e. there exists a family of reals λ_{ij}, $(i, j) \in \mathcal{I}^*$ such that*

$$\nabla l_y(\theta^*) + \sum_{(i,j)\in\mathcal{I}^*} \lambda_{ij}\nabla t_{ij}(\theta^*) = 0.$$

Proof. Let $\Phi_{ij}(\theta, \bar{\theta})$ denote the bivariate function defined by

$$\Phi_{ij}(\theta, \bar{\theta}) = \phi\Big(\frac{t_{ij}(\bar{\theta})}{t_{ij}(\theta)}\Big).$$

Let $\{\theta^{\sigma(k)}\}_{k\in\mathbb{N}}$ be a convergent subsequence of the proximal sequence with limit equal to θ^*. The first order optimality condition at iteration k is given by

$$\begin{array}{ll} \nabla l_y(\theta^{\sigma(k)}) & +\beta_{\sigma(k)}\Big(\sum_{ij}\alpha_{ij}(y_j)\nabla t_{ij}(\theta^{\sigma(k)})\phi\Big(\frac{t_{ij}(\theta^{\sigma(k)-1})}{t_{ij}(\theta^{\sigma(k)})}\Big) \\ & +\sum_{ij}\alpha_{ij}(y_j)t_{ij}(\theta^{\sigma(k)})\nabla_1\Phi(\theta^{\sigma(k)}, \theta^{\sigma(k)-1})\Big) = 0. \end{array} \tag{4.19}$$

We have

$$t_{ij}(\theta^{\sigma(k)})\nabla_1\Phi(\theta^{\sigma(k)}, \theta^{\sigma(k)-1}) = -\frac{t_{ij}(\theta^{\sigma(k)-1})}{t_{ij}(\theta^{\sigma(k)})}\phi'\Big(\frac{t_{ij}(\theta^{\sigma(k)-1})}{t_{ij}(\theta^{\sigma(k)})}\Big)\nabla t_{ij}(\theta^{\sigma(k)})$$

for all i and j.

Claim A. *For all (i, j) such that $\alpha_{ij}(y_j) \neq 0$, we have*

$$\lim_{k\to+\infty} t_{ij}(\theta^{\sigma(k)})\nabla_1\Phi(\theta^{\sigma(k)}, \theta^{\sigma(k)-1}) = 0.$$

Proof of Claim A. Two cases may occur. In the first case, we have $t_{ij}(\theta^*) = 0$. Since the sequence $\{\theta^k\}_{k\in\mathbb{N}}$ is bounded due to Lemma 3.2, continuous differentiability of ϕ and the t_{ij} proves that $\nabla_1\Phi(\theta^{\sigma(k)}, \theta^{\sigma(k)-1})$ is bounded from above. Thus, the desired conclusion follows. In the second case, $t_{ij}(\theta^*) \neq 0$ and applying Lemma 3.3, we deduce that $I_y(\theta^{\sigma(k)}, \theta^{\sigma(k)-1})$ tends to zero. Hence, $\lim_{k\to+\infty}\Phi(\theta^{\sigma(k)}, \theta^{\sigma(k)-1}) = 0$, which implies that $\lim_{k\to+\infty}\theta^{\sigma(k)}/\theta^{\sigma(k)-1} = 1$. From this and Assumptions 2.4, we deduce that $\lim_{k\to+\infty}\phi'(t_{ij}(\theta^{\sigma(k)-1})/t_{ij}(\theta^{\sigma(k)})) = 0$. Since $\{\theta^{\sigma(k)}\}_{k\in\mathbb{N}}$ converges to θ^* and that $t_{ij}(\theta^*) \neq 0$, we obtain that the subsequence $\{t_{ij}(\theta^{\sigma(k)-1})/t_{ij}(\theta^{\sigma(k)})\}_{k\in\mathbb{N}}$ is bounded from above. Moreover, $\{\nabla t_{ij}(\theta^{\sigma(k)})\}_{k\in\mathbb{N}}$ is also bounded by continuous differentiability of t_{ij}. Therefore, the fact that $\lim_{k\to+\infty}\phi'(t_{ij}(\theta^{\sigma(k)-1})/t_{ij}(\theta^{\sigma(k)})) = 0$ establishes Claim A. □

Using this claim, we just have to study the remaining right hand side terms in (3.13), namely the expression $\sum_{ij}\alpha_{ij}(y_j)\nabla t_{ij}(\theta^{\sigma(k)})\phi\Big(\frac{t_{ij}(\theta^{\sigma(k)-1})}{t_{ij}(\theta^{\sigma(k)})}\Big)$. Let $\mathcal{I}^{**}$ be a subset of the active indices $\mathcal{I}$ such that the family $\{\nabla t_{ij}(\theta^*)\}_{ij}$ is linearly independent. This linear independence is preserved under small perturbations, we may assume without loss of generality that the family $\Big\{\nabla t_{ij}(\theta^{\sigma(k)})\Big\}_{(i,j)\in\mathcal{I}^{**}}$ is linearly independent for k sufficiently large. For such k, we may rewrite equation (3.13) as

$$\begin{array}{ll} \nabla l_y(\theta^{\sigma(k)}) & +\beta_{\sigma(k)}\Big(\sum_{(i,j)\in\mathcal{I}^{**}}\lambda_{ij}^{\sigma(k)}(y_j)\nabla t_{ij}(\theta^{\sigma(k)}) \\ & +\sum_{ij}\alpha_{ij}(y_j)t_{ij}(\theta^{\sigma(k)})\nabla_1\Phi(\theta^{\sigma(k)}, \theta^{\sigma(k)-1})\Big) = 0. \end{array} \tag{4.20}$$

Claim B. *The sequence $\{\lambda_{ij}^{\sigma(k)}(y_j)\}_{k\in\mathbb{N}}$ is bounded.*

Proof of claim B. Using the previous claim and the continuous differentiability of l_y and t_{ij}, equation (3.15) expresses that $\{\lambda_{ij}^{\sigma(k)}(y_j)\}_{ij}$ are proportional to the coordinates of the projection on the span of the $\{\nabla t_{ij}(\theta^{\sigma(k)})\}_{ij}$ of a vector converging towards $\nabla l_y(\theta^*)$. Since $\{\nabla t_{ij}(\theta^{\sigma(k)})\}_{ij}$, for $(i, j) \in \mathcal{I}^{**}$, form a linearly independent family for k sufficiently large, none of the coordinates can tend towards infinity. □

We are now in position to finish the proof of the theorem. Take any cluster point τ_{ij} of $t_{ij}(\theta^{\sigma(k)-1})/t_{ij}(\theta^{\sigma(k)})$. Using Claim B, we know that $(\lambda_{ij}^{\sigma(k)}(y_j))_{(i,j)\in\mathcal{I}^{**}}$ lies in a compact set. Let $(\lambda_{ij}^*)_{(i,j)\in\mathcal{I}^{**}}$ be a cluster point of this sequence. Passing to the limit, we obtain from equation (3.13) that

$$\nabla l_y(\theta^{\sigma(k)}) + \beta^*\Big(\sum_{(i,j)\in\mathcal{I}^{**}} \lambda_{ij}^* \nabla t_{ij}(\theta^*)\Big) = 0.$$

for every cluster point β^* of $\{\beta_{\sigma(k)}\}_{k\in\mathbb{N}}$. For all $(i,j)\in\mathcal{I}^{**}$, set $\lambda_{ij} = \beta^*\lambda_{ij}^*$. This equation is exactly the Karuch-Kuhn-Tucker necessary condition for optimality. □

Remark 4.4 *If the family $(\nabla t_{ij}(\theta^{\sigma(k)}))_{(i,j)\in\mathcal{I}^*}$ is linearly independent for k sufficiently large, Theorem 3.5 holds and in addition the $\{\lambda_{ij}\}_{ij}$ are nonnegative, which proves that θ^* satisfies the Karush-Kuhn-Tucker conditions when it lies in the closure of $\mathcal{D}_I$.*

5 Application

The goal of this section is to illustrate the utility of the previous theory for a nonparametric survival analysis with competing risks proposed by Ahn, Kodell and Moon in [1].

The problem and the Kullback proximal method

This problem can be described as follows. Consider a group of N animals in an animal carcinogenecity experiment. Sacrifices are performed at certain prescribed times denoted by $t_1, t_2, \ldots, t_m$ in order to study the presence of the tumor of interest. Let T_1 be the time to onset of tumor, T_D the time to death from this tumor and X_C be the time to death from a cause other than this tumor. Notice that T_1, T_D and X_C are unobservable. The quantities to be estimated are $S(t)$, $P(t)$ and $Q(t)$, the survival function of T_1, T_D and X_C respectively. It is assumed that T_1 and T_D are statistically independent of X_C.

A nonparametric approach to estimation of S, P and Q is proposed in [1]: observed data $y_1, \ldots, y_n$ are the number of deaths on every interval $(t_j, t_{j+1}]$ which can be classified into the following four categories,

- death with tumor (without knowing cause of death)
- death without tumor
- sacrifice with tumor
- sacrifice without tumor

This gives rise to a multinomial model whose probability mass is parametrized by the values of S, P and Q at times $t_1, \ldots, t_m$. More precisely, for each time interval $(t_j, t_{j+1}]$ denote by c_j the number of deaths with tumor present, b_{1j} the number of deaths with tumor absent, a_{2j} the number of sacrifices with tumor present and b_{2j} the number of sacrifices with tumor absent. Let $N_j \leqslant N$ be the number of live animals in the population at t_j, it is shown in [1] that the corresponding log-likelihood is given by

$$\begin{aligned}\log g(y;\theta) \quad &= \textstyle\sum_{j=1}^{m}(N_{j-1}-N_j)\sum_{k=1}^{j-1}\log(p_kq_k) + (a_{2j}+b_{2j})\log(p_jq_j)\\ &+c_j\log\Big((1-p_j)+(1-\pi_jp_j)(1-q_j)\Big)\\ &+b_{1j}\log((1-q_j)\pi_{j-1}) + a_{2j}\log(1-\pi_j) + b_{2j}\log\pi_j + Cst,\end{aligned} \tag{5.21}$$

where Cst is a constant $\pi_j = S(t_j)/P(t_j)$, $p_j = P(t_j)/P(t_{j-1})$ and $q_j = Q(t_j)/Q(t_{j-1})$, $j = 1, \dots, m$, $\theta = (\pi_1, \dots, p_J, p_1, \dots, p_J, q_1, \dots, q_J)$ and the parameter space is specified by the constraints

$$\begin{aligned}\Theta = \quad & \Big\{\theta = (\pi_1, \dots, p_J, p_1, \dots, p_J, q_1, \dots, q_J) \mid 0 \leqslant \pi_j \leqslant 1, \\ & 0 \leqslant p_j \leqslant 1, \quad 0 \leqslant q_j \leqslant 1, \quad j = 1, \dots, m \text{ and } \pi_j p_j \leqslant \pi_{j-1} \quad j = 2, \dots, m\Big\},\end{aligned} \tag{5.22}$$

where the last nonconvex constraint serves to impose monotonicity of S. Note that monotonicity of P and Q is a direct consequence of the constraints on the p_j's and the q_j's, respectively.

Define the complete data $x_1, \dots, x_n$ as a measurement that indicates the cause of death in addition to the presence of absence of a tumor in the dead animals. Specifically, $x_1, \dots, x_n$ should fall into one of the following categories

- death caused by tumor
- death with incidental tumor
- death without tumor
- sacrifice with tumor
- sacrifice without tumor

To each time interval $(t_j, t_{j+1}]$ among those animals dying of natural causes, there correspond the numbers d_j of deaths caused by tumor and the number a_{1j} of deaths with incidental tumor, neither of which are observable. The associated complete log-likelihood function is given by

$$\begin{aligned}\log f(x;\theta) \quad & = \textstyle\sum_{j=1}^{m}(N_{j-1} - N_j)\sum_{k=1}^{j-1}\log(p_k q_k) + (a_{2j} + b_{2j})\log(p_j q_j) \\ & + d_j \log(1 - p_j) + a_{1j}\log\Big((1 - \pi_j p_j)(1 - q_j)\Big) \\ & + b_{1j}\log((1 - q_j)\pi_{j-1}) + a_{2j}\log(1 - \pi_j) + b_{2j}\log\pi_j + Cst\end{aligned} \tag{5.23}$$

Now, we have to compute the expectation $Q(\theta, \bar{\theta})$ of the log-likelihood function of the complete data conditionally to the parameter $\bar{\theta}$. The random variables d_j and a_{1j} are binomial with parameter λ_j and $1 - \lambda_j$ where λ_j is the probability that the death was caused by the tumor conditioned on the presence of the tumor. Conditioned on $\bar{\theta}$, we have

$$\lambda_j = \frac{1 - \bar{p}_j}{1 - \bar{p}_j + (1 - \bar{\pi}_j\bar{p}_j)(1 - \bar{q}_j)} \tag{5.24}$$

(see [1, Section 3] for details). From this, we obtain that the conditional mean values of d_j and a_{1j} are given by

$$E[d_j \mid y; \bar{\theta}] = \lambda_j c_j \quad \text{and} \quad E[a_{1j} \mid y; \bar{\theta}] = (1 - \lambda_j)c_j. \tag{5.25}$$

Therefore

$$\begin{aligned}Q(\theta, \bar{\theta}) \quad & = \textstyle\sum_{j=1}^{m}(N_{j-1} - N_j)\sum_{k=1}^{j-1}\log(p_k q_k) + (a_{2j} + b_{2j})\log(p_j q_j) \\ & + \lambda_j c_j \log(1 - p_j) + (1 - \lambda_j)c_j \log\Big((1 - \pi_j p_j)(1 - q_j)\Big) \\ & + b_{1j}\log((1 - q_j)\pi_{j-1}) + a_{2j}\log(1 - \pi_j) + b_{2j}\log\pi_j + Cst.\end{aligned} \tag{5.26}$$

From this, we can easily compute the associated Kullback distance-like function:

$$I_y(\theta,\bar\theta)=\sum_{j=1}^{m} c_j\Big(t'_j(\theta)\phi\Big(\frac{t'_j(\bar\theta)}{t'_j(\theta)}\Big)+t''_j(\theta)\phi\Big(\frac{t''_j(\bar\theta)}{t''_j(\theta)}\Big)\Big), \tag{5.27}$$

with

$$t'_j(\theta)=\frac{1-p_j}{1-p_j+(1-\pi_j p_j)(1-q_j)} \quad \text{and} \quad t''_j(\theta)=\frac{(1-\pi_j p_j)(1-q_j)}{1-p_j+(1-\pi_j p_j)(1-q_j)} \tag{5.28}$$

and ϕ is defined by $\phi(\tau)=\tau\log(\tau)$. It is straightforward to verify that Assumptions 2.2, 2.2, 2.7 and 2.4 are satisfied.

The main computational problem in this example is to handle the difficult nonconvex constraints entering the definition of the parameter space Θ. The authors of [15] and [1] use the Complex Method proposed by Box in [2] to address this problem. However, the theoretical convergence properties of Box's method are not known as reported in article MR0184734 in the Math. Reviews. Using our proximal point framework, we are able to easily incorporate the nonconvex constraints into the Kullback distance-like function and obtain an efficient algorithm with satisfactory convergence properties. For this purpose, let I'_y be defined by

$$I'_y(\theta,\bar\theta)=I_y(\theta,\bar\theta)+\sum_{j=2}^{m} t'''_j(\theta)\phi\Big(\frac{t'''_j(\bar\theta)}{t'''_j(\theta)}\Big) \tag{5.29}$$

where

$$t'''_j(\theta)=\frac{\pi_{j-1}-\pi_j p_j}{\sum_{i=2}^{m}\pi_{i-1}-\pi_i p_i}. \tag{5.30}$$

Using this new function, the nonconvex constraints $\pi_j p_j \leqslant \pi_{j-1}$ are satisfied for all proximal iterations and Assumptions 2.4 still hold.

Experimental results

We implemented the Kullback proximal algorithm with different choices of relaxation sequence $(\beta_k)_{k\in\mathbb{N}}$, $\beta_k=\beta$. The M-step of the EM algorithm does not have a closed form solution, so that nothing is lost by setting β_k to a constant not equal to one.

We attempted to supplement the KPP-EM algorithm with the Newton method and other built-in methods available in Scilab but they were not even able to find local maximizers due to the explosive nature of the logarithms near zero, leading these routines to repetitive crashes. To overcome this difficulty, we found it convenient to use the extremely simple simulated annealing random search procedure; see [22] for instance. This random search approach avoids numerical difficulties encountered using standard optimization packages and easily handles nonconvex constraints. The a.s. convergence of this procedure is well established and recent studies such as [11] confirm the good computational efficiency for convex functions optimization.

Some of our results for the data of Table 1 of [15] are given in Figures 1 to 4. In the reported experiments, we chose three constant sequences with respective values $\beta_n=$ 100, 1, .01. We observed the following phenomena

1. after one hundred iterations the increase in the likelihood function is less than 10^{-5} except for the case $\beta_n=100$ (Figure C.4) where the algorithm had not converged.

2. for $\beta_n=100$ we often obtained the best initial growth of the likelihood

3. for $\beta_n = .01$ we always obtained the highest likelihood when the number of iterations was limited to 50 (see Figure C.3 for the case MCL Male AL).

It was shown in [5] that penalizing with a parameter sequence $(\beta_n)_{n\in\mathbb{N}}$ converging towards zero implies superlinear convergence in the case where the maximum likelihood estimator lies in the interior of the constraint set. Thus, our simulations results seem to confirm observation 3. The second observation was surprising to us but this phenomenon occured repeatedly in our experiments. This behavior did not occur in our simulations for the Poisson inverse problem in [5] for instance.

In conclusion, this competing risks estimation problem is an interesting test for our Kullback-proximal method which shows that the proposed framework can provide provably convergent methods for difficult constrained nonconvex estimation problems for which standard optimization algorithms can be hard to tune. The relaxation parameter sequence $(\beta_n)_{n\in\mathbb{N}}$ also appeared crucial for this problem although the choice $\beta_n = 1$ could not really be considered unsatisfactory in practice.

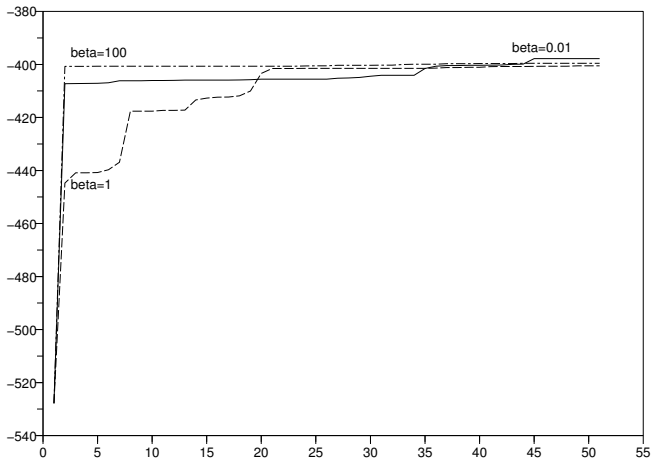

Figure C.1: Evolution of the log-likelihood versus iteration number: MCL Female CR case

6 Conclusions

The goal of this paper was the study of the asymptotic behavior of the EM algorithm and its proximal generalizations. We clarified the analysis by making use of the Kullback-proximal theoretical framework. Two of our main contributions are the following. Firstly we showed that interior cluster points are stationary points of the likelihood function and are local maximizers for sufficiently small values of β. Secondly, we showed that cluster points lying on the boundary satisfy the Karush-Kuhn-Tucker conditions. Such cases were very seldom studied in the literature although constrained estimation is a topic of growing importance; see for instance the special issue of the Journal of Statistical Planning and Inference [10] which is devoted to the problem of estimation under constraints. On the negative side, the

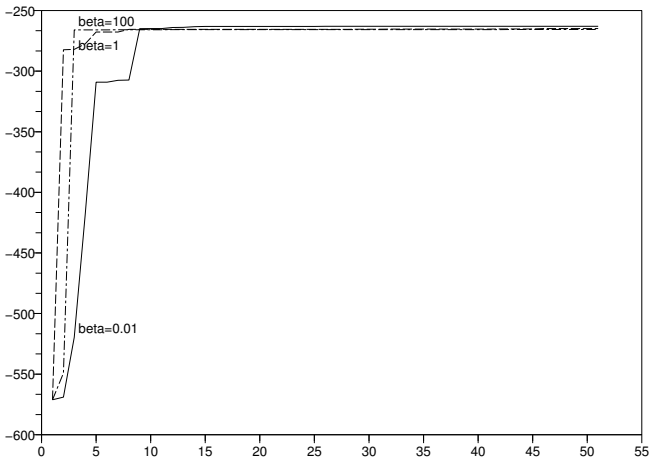

Figure C.2: Evolution of the log-likelihood versus iteration number: MCL Male AL case

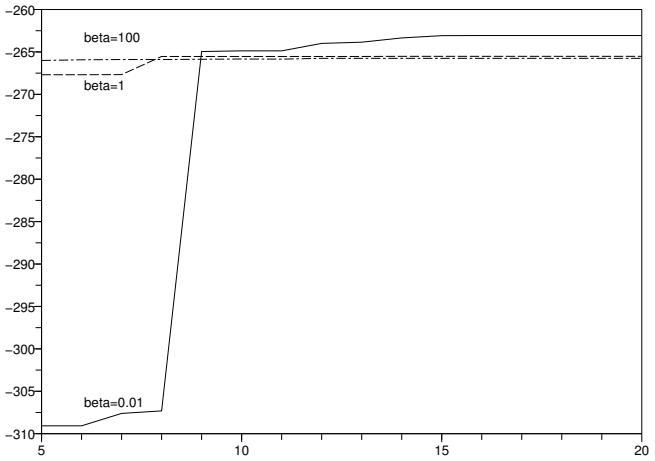

Figure C.3: Evolution of the log-likelihood versus iteration number: Detail of MCL Male AL case

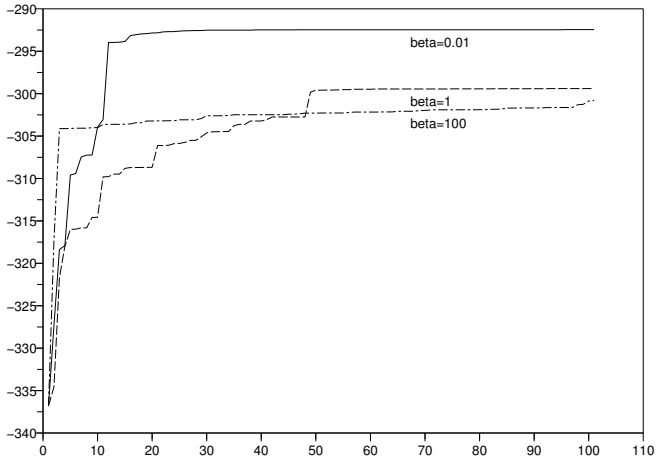

Figure C.4: Evolution of the log-likelihood versus iteration number: MCL Female AL case

analysis from the Kullback-proximal viewpoint allowed us to understand why uniqueness of the cluster point is hard to establish theoretically. On the positive side, we were able to implement a new and efficient proximal point method for estimation in the difficult tumor lethality problem involving nonlinear inequality constraints.

Bibliography

[1] H. Ahn, H. Moon and R.L. Kodell, "Attribution of tumour lethality and estimation of the time to onset of occult tumours in the absence of cause-of-death information". *J. Roy. Statist. Soc. Ser. C*, vol. 49, no. 2, 157–169, 2000. (pp. 87, 100, 101 et 102).

[2] M.J. Box, "A new method of constrained optimization and a comparison with other methods", The Computer Journal, 8, 42–52, 1965. (p. 102).

[3] G. Celeux, S. Chretien, F. Forbes and A. Mkhadri, "A component-wise EM algorithm for mixtures", J. Comput. Graph. Statist. 10 (2001), no. 4, 697–712 and INRIA RR-3746, Aug. 1999. (pp. 90 et 96).

[4] S. Chretien and A.O. Hero, "Acceleration of the EM algorithm via proximal point iterations", Proceedings of the International Symposium on Information Theory, MIT, Cambridge, p. 444, 1998. (pp. 86 et 87).

[5] S. Chrétien and A. Hero, "Kullback proximal algorithms for maximum-likelihood estimation," IEEE Trans. Inform. Theory 46 (2000), no. 5, 1800–1810. (pp. 87, 88, 89, 90, 91 et 103).

[6] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, 1987. (p. 49).

[7] I. Csiszár, "Information-type measures of divergence of probability distributions and indirect observations", Studia Sci. Math. Hung., 2 (1967), 299–318. (p. 97).

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Statistical Society, Ser. B, vol. 39, no. 1, pp. 1–38, 1977. (pp. 43, 45, 46 et 86).

[9] I. A. Ibragimov and R. Z. Has'minskii, *Statistical estimation: Asymptotic theory*, Springer-Verlag, New York, 1981. (pp. 47 et 89).

[10] Journal of Statistical Planning and Inference, vol. 107, no. 1–2, 2002. (p. 103).

[11] A. T. Kalai and S. Vempala "Simulated annealing for convex optimization". Math. Oper. Res. 31 (2006), no. 2, 253–266. (p. 102).

[12] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *Journal of Computer Assisted Tomography*, vol. 8, no. 2, pp. 306–316, 1984.

[13] B. Martinet, "Régularisation d'inéquation variationnelles par approximations successives," *Revue Francaise d'Informatique et de Recherche Operationnelle*, vol. 3, pp. 154–179, 1970. (pp. 87 et 89).

[14] G.J. McLachlan and T. Krishnan, "The EM algorithm and extensions," Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley and Sons, Inc., New York, 1997. (pp. 86 et 87).

[15] H. Moon, H. Ahn, R. Kodell and B. Pearce " A comparison of a mixture likelihood method and the EM algorithm for an estimation problme in animal carcinogenicity studies," Computational Statistics and Data Analysis, 31 , no. 2, pp. 227–238, 1999. (p. 102).

[16] A. M. Ostrowski *Solution of equations and systems of equations*. Pure and Applied Mathematics, Vol. IX. Academic Press, New York-London 1966 (p. 95).

[17] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization*, vol. 14, pp. 877–898, 1976. (pp. 87 et 89).

[18] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. on Medical Imaging*, vol. MI-1, No. 2, pp. 113–122, Oct. 1982.

[19] M. Teboulle, "Entropic proximal mappings with application to nonlinear programming," *Mathematics of Operations Research*, vol. 17, pp. 670–690, 1992. (pp. 44, 67 et 96).

[20] P. Tseng, "An analysis of the EM algorithm and entropy-like proximal point methods," *Mathematics of Operations Research*, vol. 29, pp. 27–44, 2004. (p. 90).

[21] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Annals of Statistics*, vol. 11, pp. 95–103, 1983. (pp. 86 et 95).

[22] Z. B. Zabinsky "Stochastic adaptive search for global optimization". *Nonconvex Optimization and its Applications*, 72. Kluwer Academic Publishers, Boston, MA, 2003. (p. 102).

[23] W. I. Zangwill and B. Mond, *Nonlinear programming: a unified approach*, Prentice-Hall International Series in Management. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969.

(p. 86).

Chapter D

Space Alternating Penalized Kullback Proximal Point Algorithms for Maximizing Likelihood with Nondifferentiable Penalty

with Alfred O. Hero and Hervé Perdry.

Abstract

The EM algorithm is a widely used methodology for penalized likelihood estimation. Provable monotonicity and convergence are the hallmarks of the EM algorithm and these properties are well established for smooth likelihood and smooth penalty functions. However, many relaxed versions of variable selection penalties are not smooth. In this paper we introduce a new class of Space Alternating Penalized Kullback Proximal extensions of the EM algorithm for nonsmooth likelihood inference. We show that the cluster points of the new method are stationary points even when they lie on the boundary of the parameter set. We illustrate the new class of algorithms for the problems of model selection for finite mixtures of regression and of sparse image reconstruction.

1 Introduction

The EM algorithm of Dempster Laird and Rudin (1977) is a widely applicable methodology for computing likelihood maximizers or at least stationary points. It has been extensively studied over the years and many useful generalizations have been proposed including, for instance, the stochastic EM algorithm of Delyon, Lavielle and Moulines (1999) and Kuhn and Lavielle (2004); the PX-EM accelerations of Liu, Rubin and Wu (1998); the MM generalization of Lange and Hunter (2004) and approaches using extrapolation such as proposed in Varadhan and Roland (2007).

In recent years, much attention has been given to the problem of variable selection for multiparameter estimation, for which the desired solution is sparse, i.e. many of the parameters are zero. Several approaches have been proposed for recovering sparse models. A large number of contributions are based on the use of non-differentiable penalties like the LASSO (Tibshirani (1996) and Candès and Plan (2008)), ISLE (Friedman and Popescu (2003)) and

"hidden variable"-type approach developed by Figueiredo and Nowak (2003). Other contributions are for instance sparse Bayes learning (Tipping (2001)), information theoretic based prior methods of Barron (1999), empirical Bayes (Johnstone and Silverman (2004)). Among recent alternatives is the new Dantzig selector of Candès and Tao (2008). On the other hand, only a few attempts have been made to use of non-differentiable penalization for more complex models than the linear model; for some recent progress, see Koh, Kim, and Boyd (2007) for the case of logistic regression; and Khalili and Chen (2007) for mixture models.

In the present paper, we develop new extensions of the EM algorithm that incorporate a non-differentiable penalty at each step. Following previous work of the first two authors, we use a Kullback Proximal interpretation for the EM-iterations and prove stationarity of the cluster points of the methods using nonsmooth analysis tools. Our analysis covers coordinate by coordinate methods such as Space Alternating extensions of EM and Kullback Proximal Point (KPP) methods. Such component-wise versions of EM-type algorithms can benefit from acceleration of convergence speed (Fessler and Hero (1994)). The KPP method was applied to gaussian mixture models in Celeux *et al.* (2001). The main result of this paper is that any cluster point of the Space Alternating KPP method satisfies a nonsmooth Karush-Kuhn-Tucker condition.

The paper is organized as follows. In section 2 we review Penalized Kullback Proximal Point methods and introduce componentwise PKPP algorithms with new differentiable penalties. In Section 3, our main asymptotic results are presented. In Section 4, we present a space alternating implementation of the penalized EM algorithm for a problem of model selection in a finite mixture of linear regressions using the SCAD penalty introduced in Fan and Li (2001) and further studied in Khalili and Chen (2007).

2 The EM algorithm and its Kullback proximal generalizations

The problem of maximum likelihood (ML) estimation consists of solving the maximization

$$\theta_{ML} = argmax_{\theta\in\Theta}\ l_y(\theta), \tag{2.1}$$

where y is an observed sample of a random variable Y defined on a sample space $\mathcal{Y}$ and $l_y(\theta)$ is the log-likelihood function defined by

$$l_y(\theta) = \log g(y;\theta),$$

on the parameter space $\Theta \subset \mathbb{R}^p$, and $g(y;\theta)$ denotes the density of Y at y parametrized by the vector parameter θ.

The standard EM approach to likelihood maximization introduces a complete data vector X with density f. Consider the conditional density function $k(x|y;\bar{\theta})$ of X given y

$$k(x|y;\bar{\theta}) = \frac{f(x;\bar{\theta})}{g(y;\bar{\theta})}. \tag{2.2}$$

As is well known, the EM algorithm then consists of alternating between two steps. The first step, called the E(xpectation) step, consists of computing the conditional expectation of the complete log-likelihood given Y. Notice that the conditional density k is parametrized by the current iterate of the unknown parameter value, denoted here by $\bar{\theta}$ for simplicity. Moreover, the expected complete log-likelihood is a function of the variable θ. Thus the second step, called the M(aximization) step, consists of maximizing the obtained expected complete log-likelihood with respect to the variable parameter θ. The maximizer is then

accepted as the new current iterate of the EM algorithm and the two steps are repeated until convergence is achieved.

Consider now the general problem of maximizing a concave function $\Phi(\theta)$. The original proximal point algorithm introduced by Martinet (1970) is an iterative procedure which can be written

$$\theta^{k+1} = argmax_{\theta \in D_\Phi} \left\{ \Phi(\theta) - \frac{\beta_k}{2} \|\theta - \theta^k\|^2 \right\}. \tag{2.3}$$

The influence of the quadratic penalty $\frac{1}{2}\|\theta - \theta^k\|^2$ is controlled by the sequence of positive parameters $\{\beta_k\}$. Rockafellar (1976) showed that superlinear convergence of this method occurs when the sequence $\{\beta_k\}$ converges to zero. A relationship between Proximal Point algorithms and EM algorithms was discovered in Chrétien and Hero (2000) (see also Chrétien and Hero (2008) for details). We review the EM analogy to KPP methods to motivate the space alternating generalization. Assume that the family of conditional densities $\{k(x|y;\theta)\}_{\theta \in \mathbb{R}^p}$ is regular in the sense of Ibragimov and Khasminskii (1981), in particular $k(x|y;\theta)\mu(x)$ and $k(x|y;\bar{\theta})\mu(x)$ are mutually absolutely continuous for any θ and $\bar{\theta}$ in $\mathbb{R}^p$. Then the Radon-Nikodym derivative $\frac{k(x|y,\bar{\theta})}{k(x|y;\theta)}$ exists for all $\theta, \bar{\theta}$ and we can define the following Kullback Leibler divergence:

$$I_y(\theta, \bar{\theta}) = E\big[\log \frac{k(x|y,\bar{\theta})}{k(x|y;\theta)} | y; \bar{\theta}\,\big]. \tag{2.4}$$

Let us define D_l as the domain of l_y, $D_{I,\theta}$ the domain of $I_y(\cdot,\theta)$ and D_I the domain of $I_y(\cdot,\cdot)$. Using the distance-like function I_y, the Kullback Proximal Point algorithm is defined by

$$\theta^{k+1} = argmax_{\theta \in D_\Phi} \left\{ \Phi(\theta) - \beta_k I_y(\theta, \bar{\theta}) \right\}. \tag{2.5}$$

The following was proved in Chrétien and Hero (2000).

Proposition 2.1 *[Chrétien and Hero (2000) Proposition 1]. In the case where Φ is the log-likelihood, the EM algorithm is a special instance of the Kullback-proximal algorithm with Φ equal to the penalized log-likelihood and $\beta_k = 1$, for all $k \in \mathbb{N}$.*

The Space Alternating Penalized Kullback-Proximal method

In what follows, and in anticipation of component-wise implementations of penalized KPP, we will use the notation $\Theta_r(\theta)$ for the local decomposition at θ defined by $\Theta_r(\theta) = \Theta \cap (\theta + \mathcal{S}_r)$, $r = 1, \ldots, R$ where $\mathcal{S}_1, \ldots, \mathcal{S}_R$ are subspaces of $\mathbb{R}^p$ and $\mathbb{R}^p = \oplus_{r=1}^R \mathcal{S}_r$.

Then, the Space Alternating Penalized Proximal Point Algorithm is defined as follows.

Definition 2.2 *Let ψ: $\mathbb{R}^p \mapsto \mathcal{S}_1 \times \cdots \times \mathcal{S}_R$ be a continuously differentiable mapping and let ψ_r denote its r^{th} coordinate. Let $(\beta_k)_{k\in\mathbb{N}}$ be a sequence of positive real numbers and λ be a positive real vector in $\mathbb{R}^R$. Let p_n be a nonnegative possibly nonsmooth locally Lipschitz penalty function with bounded Clarke-subdifferential (see the Appendix for details) on compact sets. Then, the Space Alternating Penalized Kullback Proximal Algorithm is defined by*

$$\theta^{k+1} = argmax_{\theta \in \Theta_{k-1(mod\ R)+1}(\theta^k) \cap D_l \cap D_{I,\theta^k}} \left\{ l_y(\theta) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta)) - \beta_k I_y(\theta, \theta^k) \right\}, \tag{2.6}$$

where D_l is the domain of l_y and $D_{I,\theta}$ is the domain of $I_y(\cdot,\theta)$.

The standard Kullback-Proximal Point algorithms as defined in Chrétien and Hero (2008) is obtained as special case by selecting $R = 1$, $\Theta_1 = \Theta$, $\lambda = 0$.

The mappings ψ_r will simply be the projection onto the subspace Θ_r, $r = 1, \ldots, R$ in the sequel but the proofs below allow for more general mappings too.

Notations and assumptions

The notation $\|\cdot\|$ will be used to denote the norm on any previously defined space. The space on which the norm operates should be obvious from the context. For any bivariate function Φ, $\nabla_1\Phi$ will denote the gradient with respect to the first variable. For the convergence analysis, we will make the following assumptions. For a locally Lipschitz function f, $\partial f(x)$ denotes the Clarke subdifferential of f at x (see the Appendix). Regular locally Lipschitz functions are defined in the Appendix.

Assumptions 2.1 *(i) l_y is differentiable and $l_y(\theta) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta))$ converges to $-\infty$ whenever $\|\theta\|$ tends to $+\infty$. The function p_n is locally Lipschitz and regular.*
(ii) The domain $D_{I,\theta}$ of $I(\cdot, \theta)$ is a subset of the domain D_l of l.
(iii) $(\beta_k)_{k\in\mathbb{N}}$ is a convergent nonnegative sequence of real numbers whose limit is denoted by β^*.
(iv) The mappings ψ_r are such that

$$\psi_r(\theta + \varepsilon d) = \psi_r(\theta)$$

for all θ in Θ, all $d \in \mathcal{S}_r^{\perp}$ and $\varepsilon > 0$ sufficiently small so that $\theta + \varepsilon d \in \Theta$, $r = 1, \ldots, R$. This condition is satisfied for linear projection operators.

We will also impose one of the two following sets of assumptions on the distance-like function I_y in (2.4).

Assumptions 2.2 *(i) There exists a finite dimensional euclidean space S, a differentiable mapping $t : D_l \mapsto S$ and a functional $\Psi : D_\Psi \subset S \times S \mapsto \mathbb{R}$ such that KL divergence (2.4) satisfies*

$$I_y(\theta, \bar{\theta}) = \Psi(t(\theta), t(\bar{\theta})),$$

where D_ψ denotes the domain of Ψ.
(ii) For any $\{(t^k, t)_{k\in\mathbb{N}}\} \subset D_\Psi$ there exists $\rho_t > 0$ such that $\lim_{\|t^k - t\|\to\infty} I_y(t^k, t) \geqslant \rho_t$. Moreover, we assume that $\inf_{t\in M} \rho_t > 0$ for any bounded set $M \subset S$.
For all (t', t) in D_Ψ, we will also require that
(iii) (Positivity) $\Psi(t', t) \geqslant 0$,
(iv) (Identifiability) $\Psi(t', t) = 0 \Leftrightarrow t = t'$,
(v) (Continuity) Ψ is continuous at (t', t)
and for all t belonging to the projection of D_Ψ onto its second coordinate,
(vi) (Differentiability) the function $\Psi(\cdot, t)$ is differentiable at t.

In the case where the Kullback divergence I_y is not defined everywhere (for instance if its domain of definition is the positive orthant), we need stronger assumptions to prove the desired convergence properties.

Assumptions 2.3 *(i) There exists a differentiable mapping $t : D_l \mapsto \mathbb{R}^{n\times m}$ such that the Kullback distance-like function I_y is of the form*

$$I_y(\theta, \bar{\theta}) = \sum_{1\leqslant i\leqslant n, 1\leqslant j\leqslant m} \alpha_{ij}(y_j) t_{ij}(\theta) \phi\Big(\frac{t_{ij}(\bar{\theta})}{t_{ij}(\theta)}\Big),$$

where for all i and j, t_{ij} is continuously differentiable on its domain of definition, α_{ij} is a function from $\mathcal{Y}$ to $\mathbb{R}_+$, the set of positive real numbers,
(ii) The function ϕ is a non negative differentiable convex function defined $\mathbb{R}_*^+$ and such that $\phi(\tau) = 0$ if and only if $\tau = 1$.
(iii) There exists $\rho > 0$ such that

$$\lim_{\mathbb{R}_+ \ni \tau \to \infty} \phi(\tau) \geqslant \rho.$$

(iv) The mapping t is injective on each Θ_r.

In the context of Assumptions 2.7, D_I is simply the set

$$D_I = \{\theta \in \mathbb{R}^p \mid t_{ij}(\theta) > 0 \quad \forall i \in \{1, \ldots, n\} \text{ and } j \in \{1, \ldots, m\}\}^2.$$

Notice that if $t_{ij}(\theta) = \theta_i$ and $\alpha_{ij} = 1$ for all i and all j, the functions I_y turn out to reduce to the well known ϕ divergence defined in Csiszàr (1967). Assumptions 2.7 are satisfied by most standard examples (for instance Gaussian mixtures and Poisson inverse problems) with the choice $\phi(\tau) = \tau \log(\tau) - 1$.

Assumptions 2.2(i) and (ii) on l_y are standard and are easily checked in practical examples, e.g. they are satisfied for the Poisson and additive mixture models.

Finally we make the following general assumption.

Assumptions 2.4 *The Kullback proximal iteration (2.6) is well defined, i.e. there exists at least one maximizer of (2.6) at each iteration k.*

In the EM case, i.e. $\beta = 1$, this last assumption is equivalent to the computability of M-steps. In practice it suffices to show the inclusion $0 \in \nabla l_y(\theta) - \lambda \partial p_n(\psi(\theta)) - \beta_k \nabla I_y(\theta, \theta^k)$ for $\theta = \theta^{k+1}$ in order to prove that the solution is unique. Then assumption 2.2(i) is sufficient for a maximizer to exist.

These technical assumptions play an important role in the theory developed below. Assumption 1 (i) on differentiability of the log-likelihood is important for establishing the Karush-Kuhn-Tucker optimality conditions for cluster points. The fact that the objective should decrease to negative infinity as the norm of the parameter goes to infinity is often satisfied, or can be easily imposed, and is used later to garantee boundedness of the sequence of iterates. The fact that p_n is regular is standard since the usual choices are the ℓ_1-norm, the ℓ_p-quasi-norms for $0 < p < 1$, the SCAD penalty, etc ... Assumption 1 (ii) is only needed in order to simplify the analysis since, otherwise, each iterate would lie in the intersection of D_l and D_I and this would lead to asymptotic complications; this assumption is always satisfied in the models we have encountered in practice. Assumption 1 (iii) is standard. Assumption 1 (iv) is satisfied when ψ_r is a projection onto $\mathcal{S}_r$ and simplifies the proofs. Assumption 2 imposes natural conditions on the "distance" I_y. Assumption 2 (ii) ensures that the "distance" I_y is large between points whose euclidean distance goes to $+\infty$, thus weakening the assumption that I_y should grow to $+\infty$ in such a case. Assumptions 3 are used to obtain the Karush-Kuhn-Tucker conditions in Theorem 2. For this Theorem, we require I_y to behave like a standard Kullback-Leibler "distance" and therefore that I_y has a more constrained shape. Assumption 3 (iii) is a simplification of Assumption 2 (ii). Assumption 3 (iv) is a natural injectivity requirement.

3 Asymptotic properties of the Kullback-Proximal iterations

Basic properties of the penalized Kullback proximal algorithm

Under Assumptions 2.2, we state basic properties of the penalized Kullback Proximal Point Algorithm. The most basic property is the monotonicity of the penalized likelihood function and the boundedness of the penalized proximal sequence $(\theta^k)_{k\in\mathbb{N}}$. The proofs of the following lemmas are given, for instance, in Chrétien and Hero (2000) for the unpenalized case ($\lambda = 0$) and their generalizations to the present context is straightforward.

We start with the following monotonicity result.

Lemma 3.1 *For any iteration $k \in \mathbb{N}$, the sequence $(\theta^k)_{k\in\mathbb{N}}$ satisfies*

$$l_y(\theta^{k+1}) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta^{k+1})) - (l_y(\theta^k) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta^k))) \geqslant \beta_k I_y(\theta^k, \theta^{k+1}) \geqslant 0. \quad (3.7)$$

Lemma 3.2 *The sequence $(\theta^k)_{k\in\mathbb{N}}$ is bounded.*

The next lemma will also be useful and its proof in the unpenalized case where $\lambda = 0$ is given in Chrétien and Hero (2008) Lemma 2.4.3. The generalization to $\lambda > 0$ is also straightforward.

Lemma 3.3 *Assume that in the Space Alternating KPP sequence $(\theta^k)_{k\in\mathbb{N}}$, there exists a subsequence $(\theta^{\sigma(k)})_{k\in\mathbb{N}}$ belonging to a compact set C included in D_l. Then,*

$$\lim_{k\to\infty} \beta_k I_y(\theta^{k+1}, \theta^k) = 0.$$

One important property, which is satisfied in practice, is that the distance between two successive iterates decreases to zero. This property is critical to the definition of a stopping rule for the algorithm. This property was established in Chrétien and Hero (2008) in the case $\lambda = 0$.

Proposition 3.4 *[Chrétien and Hero (2008) Proposition 4.1.2] The following statements hold.*

(i) For any sequence $(\theta^k)_{k\in\mathbb{N}}$ in $\mathbb{R}_+^p$ and any bounded sequence $(\eta^k)_{k\in\mathbb{N}}$ in $\mathbb{R}_+^p$, if $\lim_{k\to+\infty} I_y(\eta^k, \theta^k) = 0$ then $\lim_{k\to+\infty} |t_{ij}(\eta^k) - t_{ij}(\theta^k)| = 0$ for all i,j such that $\alpha_{ij} \neq 0$.

(ii) If $\lim_{k\to+\infty} I_y(\eta^k, \theta^k) = 0$ and one coordinate of one of the two sequences $(\theta^k)_{k\in\mathbb{N}}$ and $(\eta^k)_{k\in\mathbb{N}}$ tends to infinity, so does the other's same coordinate.

Properties of cluster points

The results of this subsection state that any cluster point θ^* such that (θ^*, θ^*) lies on the closure of D_I satisfies a modified Karush-Kuhn-Tucker type condition. We first establish this result in the case where Assumptions 2.2 hold in addition to Assumptions 2.2 and 2.2 for the Kullback distance-like function I_y.

For notational convenience, we define

$$F_\beta(\theta, \bar{\theta}) = l_y(\theta) - \sum_{r=1}^{R} \lambda_r p_n(\psi_r(\theta)) - \beta I_y(\theta, \bar{\theta}). \quad (3.8)$$

Theorem 3.5 *Assume that Assumptions 2.2, 2.2 and 2.4 hold and if $R > 1$, then, for each $r = 1, \ldots, R$, t is injective on Θ_r. Assume that the limit of $(\beta_k)_{k\in\mathbb{N}}$, β^*, is positive. Let θ^* be a cluster point of the Space Alternating Penalized Kullback-proximal sequence (2.6). Assume the mapping t is differentiable at θ^*. If θ^* lies in the interior of D_l, then θ^* is a stationary point of the penalized log-likelihod function $l_y(\theta)$, i.e.*

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^{R} \lambda_r \partial p_n(\psi_r(\theta^*)).$$

Proof. We consider two cases, namely the case where $R = 1$ and the case where $R > 1$.

A. If $R = 1$ the proof is analogous to the proof of Theorem 3.2.1 in Chrétien and Hero (2008). In particular, we have

$$F_{\beta^*}(\theta^*, \theta^*) \geqslant F_{\beta^*}(\theta, \theta^*)$$

for all θ such that $(\theta, \theta^*) \in D_I$. Since $I_y(\theta, \theta^*)$ is differentiable at θ^*, the result follows by writing the first order optimality condition at θ^* in (1.1).

B. Assume that $R > 1$ and let $(x^{\sigma(k)})_{k\in\mathbb{N}}$ be a subsequence of iterates of (2.6) converging to θ^*. Moreover let $r = 1, \ldots, R$ and $\theta \in \Theta_r \cap D_l$. For each k, let $\sigma_r(k)$ the smallest index greater than $\sigma(k)$, of the form $\sigma(k') - 1$, with $k' \in \mathbb{N}$ and $(\sigma(k') - 1)\,(mod\, R) + 1 = r$. Using the fact that t is injective on every Θ_r, $r = 1, \ldots, R$, Lemma 3.3 and the fact that $(\beta_k)_{k\in\mathbb{N}}$ converges to $\beta^* > 0$, we easily conclude that $(\theta^{\sigma_r(k)})_{k\in\mathbb{N}}$ and $(\theta^{\sigma_r(k)+1})_{k\in\mathbb{N}}$ also converge to θ^*.

For k sufficiently large, we may assume that the terms $(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)})$ and $(\theta, \theta^{\sigma_r(k)})$ belong to a compact neighborhood C^* of (θ^*, θ^*) included in D_I. By Definition 2.2 of the Space Alternating Penalized Kullback Proximal iterations,

$$F_{\beta_{\sigma_r(k)}}(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}) \geqslant F_{\beta_{\sigma_r(k)}}(\theta, \theta^{\sigma_r(k)}).$$

Therefore,

$$\begin{aligned} F_{\beta^*}(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}) \quad & -(\beta_{\sigma_r(k)} - \beta^*) I_y(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}) \geqslant \\ & F_{\beta^*}(\theta, \theta^{\sigma_r(k)}) - (\beta_{\sigma_r(k)} - \beta^*) I_y(\theta, \theta^{\sigma(k)}). \end{aligned} \tag{3.9}$$

Continuity of F_β follows directly from the proof of Theorem 3.2.1 in Chrétien and Hero (2008), where in that proof $\sigma(k)$ has to be replaced by $\sigma_r(k)$. This implies that

$$F_{\beta^*}(\theta^*, \theta^*) \geqslant F_{\beta^*}(\theta, \theta^*) \tag{3.10}$$

for all $\theta \in \Theta_r$ such that $(\theta, \theta^*) \in C^* \cap D_I$. Finally, recall that no assumption was made on θ, and that C^* is a compact neighborhood of θ^*. Thus, using the assumption 2.2(i), which asserts that $l_y(\theta)$ tends to $-\infty$ as $\|\theta\|$ tends to $+\infty$, we may deduce that (3.10) holds for any $\theta \in \Theta_r$ such that $(\theta, \theta^*) \in D_I$ and, letting ε tend to zero, we see that θ^* maximizes $F_{\beta^*}(\theta, \theta^*)$ for all $\theta \in \Theta_r$ such that (θ, θ^*) belongs to D_I as claimed.

To conclude the proof of Theorem 3.5, take d in $\mathbb{R}^p$ and decompose d as $d = d_1 + \cdots + d_R$ with $d_r \in \mathcal{S}_r$. Then, equation (3.10) implies that the directional derivatives satisfy

$$F'_{\beta^*}(\theta^*, \theta^*; d_r) \leqslant 0 \tag{3.11}$$

for all $r = 1, \ldots, R$. Due to Assumption 2.2 (iv), the directional derivative of $\sum_{r=1}^{R} \lambda_r p_n(\psi_r(\cdot))$ in the direction d is equal to the sum of the partial derivatives in the directions $d_1, \ldots, d_R$

and, since all other terms in the definition of F_β are differentiable, we obtain using (3.11), that

$$F'_{\beta^*}(\theta^*, \theta^*; d) = \sum_{r=1}^{R} F'_{\beta^*}(\theta^*, \theta^*; d_r) \leqslant 0.$$

Therefore, using the assumption that p_n is regular (see Asssumption 1(i)) which says that $p_n^\circ = p'_n$, together with characterization (6.22) of the subdifferential in the Appendix and Proposition 2.1.5 (a) in [Clarke (1990)], the desired result follows. □

Next, we consider the case where Assumptions 2.7 hold.

Theorem 3.6 *Assume that in addition to Assumptions 2.2 and 2.4, Assumptions 2.7 hold. Let θ^* be a cluster point of the Space Alternating Penalized Kullback Proximal sequence. Assume that all the functions t_{ij} are continuously differentiable at θ^*. Let $\mathcal{I}^*$ denote the index of the active constraints at θ^*, i.e. $\mathcal{I}^* = \{(i,j) \text{ s.t. } t_{ij}(\theta^*) = 0\}$. If θ^* lies in the interior of D_l, then θ^* satisfies the following property: there exists a family of subsets $\mathcal{I}_r^{**} \subset \mathcal{I}^*$ and a set of real numbers λ_{ij}^*, $(i,j) \in \mathcal{I}_r^{**}$, $r = 1, \ldots, R$ such that*

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^{R} \lambda_r \partial p_n(\psi_r(\theta^*)) + \sum_{r=1}^{R} \sum_{(i,j)\in\mathcal{I}_r^{**}} \lambda_{ij}^* P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)), \tag{3.12}$$

where $P_{\mathcal{S}_r}$ is the projection onto $\mathcal{S}_r$.

Remark 3.7 *The condition (3.12) resembles the traditional Karush-Kuhn-Tucker conditions of optimality but is in fact weaker since the vector*

$$\sum_{r=1}^{R} \sum_{(i,j)\in\mathcal{I}_r^{**}} \lambda_{ij}^* P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))$$

in equation (3.12) does not necessarily belong to the normal cone at θ^* to the set $\{\theta \mid t_{ij} \geqslant 0,\ i = 1, \ldots, n,\ j = 1, \ldots, m\}$.

Proof of Theorem 3.6. Let $\Phi_{ij}(\theta, \bar{\theta})$ denote the bivariate function defined by

$$\Phi_{ij}(\theta, \bar{\theta}) = \phi\Big(\frac{t_{ij}(\bar{\theta})}{t_{ij}(\theta)}\Big).$$

As in the proof of Theorem 3.5, let $(x^{\sigma(k)})_{k\in\mathbb{N}}$ be a subsequence of iterates of (2.6) converging to θ^*. Moreover let $r = 1, \ldots, R$ and $\theta \in \Theta_r \cap D_l$. For each k, let $\sigma_r(k)$ be the next index greater than $\sigma(k)$ such that $(\sigma_r(k) - 1)\ (mod\ R) + 1 = r$. Using the fact that t is injective on every Θ_r, $r = 1, \ldots, R$, Lemma 3.3 and the fact that $(\beta_k)_{k\in\mathbb{N}}$ converges to $\beta^* > 0$, we easily conclude that $(\theta^{\sigma_r(k)})_{k\in\mathbb{N}}$ and $(\theta^{\sigma_r(k)+1})_{k\in\mathbb{N}}$ also converge to θ^*.

Due to Assumption 2.7 (iv), the first order optimality condition at iteration $\sigma_r(k)$ can be written

$$\begin{aligned} 0 = \quad & P_{\mathcal{S}_r}(\nabla l_y(\theta^{\sigma(k)+1})) - \lambda_r g_r^{\sigma_r(k)+1} + \beta_{\sigma_r(k)} \Big(\textstyle\sum_{ij} \alpha_{ij}(y_j) P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^{\sigma_r(k)+1})) \\ & \Phi_{ij}(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)}) + \textstyle\sum_{ij} \alpha_{ij}(y_j) t_{ij}(\theta^{\sigma_r(k)+1}) P_{\mathcal{S}_r}(\nabla_1 \Phi_{ij}(\theta^{\sigma_r(k)+1}, \theta^{\sigma_r(k)})) \Big) \end{aligned} \tag{3.13}$$

with $g_r^{\sigma_r(k)+1} \in \partial p_n(\psi_r(\theta^{\sigma_r(k)+1}))$.

Moreover, Claim A in the proof of Theorem 4.2.1 in Chrétien and Hero (2008), gives that for all (i,j) such that $\alpha_{ij}(y_j) \neq 0$

$$\lim_{k\to+\infty} t_{ij}(\theta^{\sigma_r(k)+1})\nabla_1\Phi_{ij}(\theta^{\sigma_r(k)+1},\theta^{\sigma_r(k)}) = 0. \tag{3.14}$$

Let $\mathcal{I}_r^*$ be a subset of indices such that the family $\{P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))\}_{(i,j)\in\mathcal{I}_r^*}$ is linearly independent and spans the linear space generated by the family of all projected gradients $\{P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))\}_{i=1,\ldots,n,j=1,\ldots,m}$. Since this linear independence are preserved under small perturbations (continuity of the gradients), we may assume, without loss of generality, that the family

$$\Big\{P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^{\sigma_r(k)+1}))\Big\}_{(i,j)\in\mathcal{I}_r^*}$$

is linearly independent for k sufficiently large. For such k, we may thus rewrite equation (3.13) as

$$\begin{aligned} 0 = \quad & P_{\mathcal{S}_r}(\nabla l_y(\theta^{\sigma_r(k)+1})) - \lambda_r g_r^{\sigma_r(k)+1} + \beta_{\sigma_r(k)}\Big(\textstyle\sum_{(i,j)\in\mathcal{I}_r^*}\pi_{ij}^{\sigma_r(k)+1}(y_j) \\ & P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^{\sigma_r(k)+1})) + \textstyle\sum_{ij}\alpha_{ij}(y_j)t_{ij}(\theta^{\sigma_r(k)+1})P_{\mathcal{S}_r}(\nabla_1\Phi(\theta^{\sigma_r(k)+1},\theta^{\sigma_r(k)}))\Big), \end{aligned} \tag{3.15}$$

where

$$\pi_{ij}^{\sigma_r(k)+1}(y_j) = \alpha_{ij}(y_j)\Phi_{ij}(\theta^{\sigma_r(k)+1},\theta^{\sigma_r(k)}). \tag{3.16}$$

Claim. *The sequence* $\{\pi_{ij}^{\sigma_r(k)+1}(y_j)\}_{k\in\mathbb{N}}$ *has a convergent subsequence for all* (i,j) *in* I_r^*.

Proof of the claim. Since the sequence $(\theta^k)_{k\in\mathbb{N}}$ is bounded, ψ is continuously differentiable and the penalty p_n has bounded subdifferential on compact sets, there exists a convergent subsequence $(g_r^{\sigma_r(\gamma(k))+1})_{k\in\mathbb{N}}$ with limit g_r^*. Now, using Equation (3.14), this last equation implies that $\{\pi_{(i,j)\in\mathcal{I}_r^*}^{\sigma_r(\gamma(k))+1}(y_j)\}_{(i,j)\in\mathcal{I}_r^*}$ converges to the coordinates of a vector in the linearly independent family $\{P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))\}_{(i,j)\in\mathcal{I}_r^*}$. This concludes the proof. □

The above claim allows us to finish the proof of Theorem 3.6. Since a subsequence $(\pi_{ij}^{\sigma_r(\gamma(k))+1}(y_j))_{(i,j)\in\mathcal{I}_r^*}$ is convergent, we may consider its limit $(\pi_{ij}^*)_{(i,j)\in\mathcal{I}_r^*}$. Passing to the limit, we obtain from equation (3.13) that

$$0 = P_{\mathcal{S}_r}(\nabla l_y(\theta^*)) - \lambda_r g_r^* + \beta^*\Big(\sum_{(i,j)\in\mathcal{I}_r^*}\pi_{ij}^* P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))\Big). \tag{3.17}$$

Using the outer semi-continuity property of the subdifferential of locally Lipschitz functions (see Appendix) we thus obtain that $g_r^* \in \partial p_n(\psi_r(\theta^*))$. Now, summing over r in (3.17), we obtain

$$0 = \sum_{r=1}^R P_{\mathcal{S}_r}(\nabla l_y(\theta^*)) - \sum_{r=1}^R \lambda_r g_r^* + \beta^*\sum_{r=1}^R\Big(\sum_{(i,j)\in\mathcal{I}_r^*}\pi_{ij}^* P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))\Big).$$

Moreover, since $\Phi_{ij}(\theta^{\sigma_r(k)+1},\theta^{\sigma_r(k)})$ tends to zero if $(i,j)\notin\mathcal{I}^*$, i.e. if the constraint on component (i,j) is not active, equation (3.16) implies that

$$0 = \sum_{r=1}^R P_{\mathcal{S}_r}(\nabla l_y(\theta^*)) - \sum_{r=1}^R \lambda_r g_r^* + \beta^*\sum_{r=1}^R\Big(\sum_{(i,j)\in\mathcal{I}_r^{**}}\pi_{ij}^* P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*))\Big)$$

where $\mathcal{I}_r^{**}$ is the subset of active indices of $\mathcal{I}_r^*$, i.e. $\mathcal{I}_r^{**} = \mathcal{I}_r^* \cap \mathcal{I}^*$. Since $\sum_{r=1}^R \lambda_r g_r^* \in \sum_{r=1}^R \lambda_r \partial p_n(\psi_r(\theta^*))$, this implies that

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^R \lambda_r \partial p_n(\psi_r(\theta^*)) + \beta^* \sum_{r=1}^R \sum_{(i,j)\in\mathcal{I}_r^{**}} \pi_{ij}^* P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)), \tag{3.18}$$

which establishes Theorem 3.6 once we define $\lambda_{ij}^* = \lambda^* \pi_{ij}^*$. □

The result (3.18) can be refined to the classical Karush-Kuhn-Tucker type condition under additional conditions such as stated below.

Corollary 3.8 *If in addition to the assumptions of Theorem 3.6 we assume that either* $P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)) = \nabla t_{ij}(\theta^*)$ *or* $P_{\mathcal{S}_r}(\nabla t_{ij}(\theta^*)) = 0$ *for all* $(i,j) \in \mathcal{I}^*$, *i.e. such that* $t_{ij}(\theta^*) = 0$, *then there exists a set of subsets* $\mathcal{I}_r^{**} \subset \mathcal{I}^*$ *and a family of real numbers* λ_{ij}^*, $(i,j) \in \mathcal{I}_r^{**}$, $r = 1, \ldots, R$ *such that the following Karush-Kuhn-Tucker condition for optimality holds at cluster point* θ^**:*

$$0 \in \nabla l_y(\theta^*) - \sum_{r=1}^R \lambda_r \partial p_n(\psi_r(\theta^*)) + \sum_{r=1}^R \sum_{(i,j)\in\mathcal{I}_r^{**}} \lambda_{ij}^* \nabla t_{ij}(\theta^*).$$

4 Application: Variable selection in finite mixtures of regression models

Variable subset selection in regression models is frequently performed using penalization of the likelihood function, e.g. using AIC, Akaike (1973) and BIC, Schwarz (1978) penalties. The main drawback of these approaches is lack of scalability due to a combinatorial explosion of the set of possible models as the number of variables increases. Newer methods use l_1-type penalties of likelihood functions, as in the LASSO, Tibshirani (1996) and the Dantzig selector of Candès and Tao (2007), to select subsets of variables without enumeration.

Computation of maximizers of the penalized likelihood function can be performed using standard algorithms for nondifferentiable optimization such as bundle methods, as introduced in Hiriart-Urruty and Lemaréchal (1993). However general purpose optimization methods might be difficult to implement in the situation where, for instance, log objective functions induce line-search problems. In certain cases, the EM algorithm, or a combination of EM type methods with general purpose optimization routines might be simpler to implement. Variable selection in finite mixture models, as described in Khalili and Chen (2007), represents such a case due to the presence of very natural hidden variables.

In the finite mixture estimation problem considered here, $y_1, \ldots, y_n$ are realizations of the response variable Y and $x_1, \ldots, x_n$ are the associated realizations of the P-dimensional vector of covariates X. We focus on the case of a mixture of linear regression models sharing the same variance, as in the baseball data example of section 7.2 in Khalili and Chen (2007), i.e.

$$Y \sim \sum_{k=1}^K \pi_k \mathcal{N}(X^t \beta_k, \sigma^2), \tag{4.19}$$

with $\pi_1, \ldots, \pi_k \geqslant 0$ and $\sum_{k=1}^K \pi_k = 1$. The main problem discussed in Khalili and Chen (2007) is model selection for which a generalization of the smoothly clipped absolute deviation (SCAD) method of Fan and Li (2001,2002) is proposed using an MM-EM algorithm in the spirit of Hunter and Lange (2004). No convergence property of the MM algorithm was established. The purpose of this section is to show that the Space Alternating KPP EM

generalization is easily implemented and that stationarity of the cluster points is garanteed by the theoretical analysis of Section 3.

The SCAD penalty, studied in Khalili and Chen (2007) is a modification of the l_1 penalty which is given by

$$p_n(\beta_1, \ldots, \beta_K) = \sum_{k=1}^{K} \pi_k \sum_{j=1}^{P} p_{\gamma_{nk}}(\beta_{k,j})$$

where p_{nk} is specified by

$$p'_{\gamma_{nk}}(\beta) = \gamma_{nk}\sqrt{n} 1_{\sqrt{n}|\beta| \leqslant \gamma_{nk}} + \frac{\sqrt{n}(a\gamma_{nk} - \sqrt{n}|\beta|)_+}{a-1} 1_{\sqrt{n}|\beta| > \gamma_{nk}}$$

for β in $\mathbb{R}$.

Define the missing data as the class labels $z_1, \ldots, z_n$ of the mixture component from which the observed data point y_n was drawn. The complete log-likelihood is then

$$l_c(\beta_1, \ldots, \beta_K, \sigma^2) = \sum_{i=1}^{n} \log(\pi_{z_i}) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - x_i^t \beta_{z_i})^2}{2\sigma^2}.$$

Setting $\theta = (\pi_1, \ldots, \pi_K, \beta_1, \ldots, \beta_K, \sigma^2)$, the penalized Q-function is given by

$$Q(\theta, \bar{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}(\bar{\theta}) \left[\log(\pi_k) - \frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - x_i^t \beta_k)^2}{2\sigma^2} \right] - p_n(\beta_1, \ldots, \beta_K)$$

where

$$t_{ik}(\theta) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y_i - X\beta_k)^2}{2\sigma^2} \right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y_i - X\beta_l)^2}{2\sigma^2} \right)}.$$

The computation of this Q-function accomplishes the E-step. Moreover, a penalty of the form $-\sum_{k=1}^{K}\sum_{j=1}^{P} |\max\{10^6, |\beta_{k,j}|\} - 10^6|$ can be added to the log-likelihood function in order to ensure that Assumptions 1(i) (convergence of the penalized log-likelihood to $-\infty$ for parameter values with norm growing to $+\infty$) is satisfied for the case where X is not invertible. Due to the fact that the penalty p_n is a function of the mixture probabilities π_k, the M-step estimate of the π vector is not given by the usual formula

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} t_{ik}(\bar{\theta}) \quad k = 1, \ldots, K. \tag{4.20}$$

This, however, is the choice made in Khalili and Chen (2007) in their implementation. Moreover, optimizing jointly over the variables β_k and π_k is clearly a more complicated task than independently optimizing with respect to each variable. We implement a componentwise version of EM consisting of successively optimizing with respect to the π_k's and alternatively with respect to the vectors β_k. Optimization with respect to the π_k's can be easily performed using standard differentiable optimization routines and optimization with respect to the β_k's can be performed by a standard non-differentiable optimization routine, e.g. as provided by the function optim of Scilab using the 'nd' (standing for 'non-differentiable') option.

We now turn to the description of the Kullback proximal penalty I_y defined by (2.4). The conditional density function $k(y_1, \ldots, y_n, z_1, \ldots, z_n \mid y_1, \ldots, y_n; \theta)$ is

$$k(y_1, \ldots, y_n, z_1, \ldots, z_n \mid y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} t_{iz_i}(\theta).$$

and therefore, the Kullback distance-like function $I_y(\theta, \bar{\theta})$ is

$$I_y(\theta, \bar{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}(\bar{\theta}) \log \Big(\frac{t_{ik}(\bar{\theta})}{t_{ik}(\theta)} \Big). \tag{4.21}$$

We have $R = K + 1$ subsets of variables with respect to which optimization will be performed successively. All components of Assumptions 2.2 and 2.7 are trivially satisfied for this model. Validation of Assumption 2.7 (iv) is provided by Lemma 1 of Celeux *et al.* (2001). On the other hand, since $t_{ik}(\theta) = 0$ implies that $\pi_k = 0$ and $\pi_k = 0$ implies

$$\frac{\partial t_{ik}}{\partial \beta_{jl}}(\theta) = 0$$

for all $j = 1, \ldots, p$ and $l = 1, \ldots, K$ and

$$\frac{\partial t_{ik}}{\partial \sigma^2}(\theta) = 0,$$

it follows that $P_{\mathcal{S}_r}(\nabla t_{ik}(\theta^*)) = \nabla t_{ik}(\theta^*)$ if $\mathcal{S}_r$ is the vector space generated by the probability vectors π and $P_{\mathcal{S}_r}(\nabla t_{ik}(\theta^*)) = 0$ otherwise. Therefore, Corollary 3.8 applies.

We illustrate this algorithm on real data (available at *http://www.amstat.org/publications/jse/v6n2/datasets.watnik.html*).

Khalili and Chen (2007) report that a model with only two components was selected by the BIC criterion in comparison to a three components model. Here, two alternative algorithms are compared: the approximate EM using (4.20) and the plain EM using the optim subroutines. The results for $\gamma_{nk} = 1$ and $a = 10$ are given in Figures D.1.

The results shown in Figure D.1 establish that the approximate EM algorithm has similar properties to the plain EM algorithm for small values of the threshold parameters γ_{nk}. Moreover, the larger the values of γ_{nk}, the closer the probability of the first component is to 1. One important fact to notice is that with the plain EM algorithm, the optimal probability vector becomes singular, in the sense that the second component has zero probability, as shown in Figure D.2 . Figure D.3 demonstrates that the approximate EM algorithm of Khalili and Chen (2007) does not produce optimal solutions.

5 Conclusion

In this paper we analyzed the expectation maximization (EM) algorithm with non-differentiable penalty. By casting the EM algorithm as a Kullback Proximal Penalized (KPP) iteration, we proved the stationarity of the cluster points and showed that any cluster point of the Space Alternating KPP method satisfies a nonsmooth Karush-Kuhn-Tucker condition. The theory was applied to a space alternating implementation of the penalized EM algorithm for a problem of model selection in a finite mixture of linear regressions.

6 Appendix: The Clarke subdifferential of a locally Lipschitz function

Since we are dealing with non differentiable functions, the notion of generalized differentiability is required. The main references for this appendix are Clarke (1990) and Rockafellar and Wets (2004). A locally Lipschitz function f: $\mathbb{R}^p \mapsto \mathbb{R}$ always has a generalized directional derivative $f^\circ(\theta, \omega)$: $\mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ in the sense given by Clarke, i.e.

$$f^\circ(\theta, \omega) = \lim \sup_{\eta \in \mathbb{R}^p \to \theta,\, t \downarrow 0} \frac{f(\eta + t\omega) - f(\eta)}{t}.$$

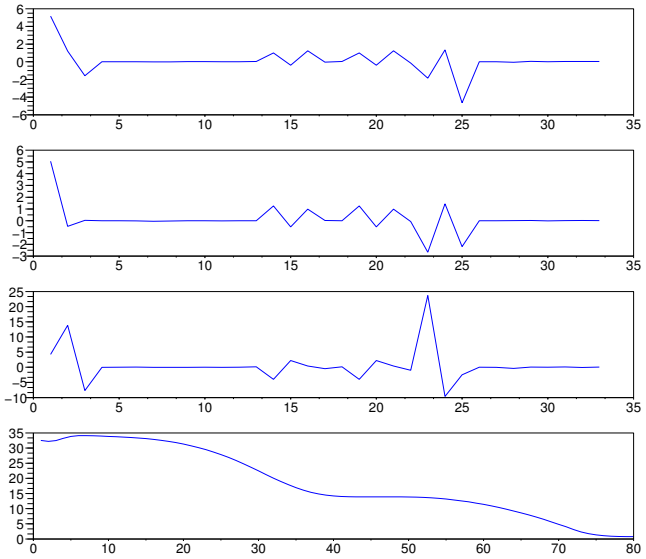

Figure D.1: Baseball data of Khalili and Chen (2007). This experiment is performed with the plain EM. The parameters are $\gamma_{nk} = .1$ and $a = 10$. The first plot is the vector β obtained for the single component model. The second (resp. third) plot is the vector of the optimal β_1 (resp. β_2). The fourth plot is the euclidean distance to the optimal θ^* versus iteration index. The starting value of π_1 was .3

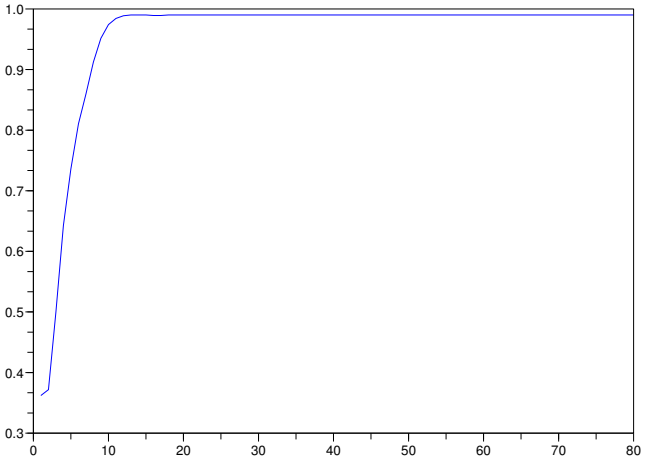

Figure D.2: This experiment is performed with the plain EM for the Baseball data of Khalili and Chen (2007). The parameters are $\gamma_{nk} = 5$ and $a = 10$. The plot shows the probability π_1 of the first component versus iteration index. The starting value of π_1 was .3

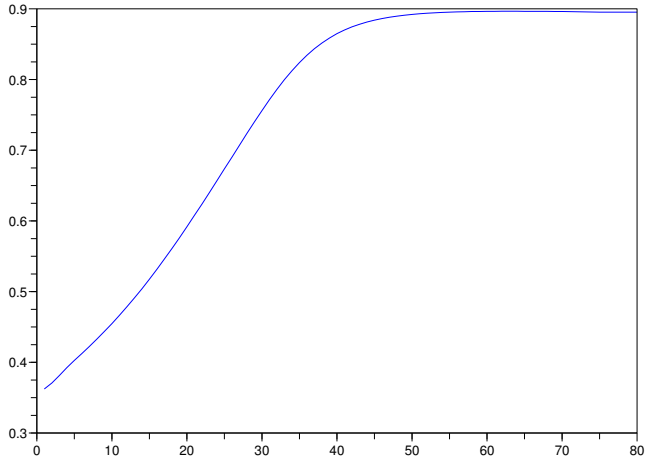

Figure D.3: Baseball data of Khalili and Chen (2007). This experiment is performed with the approximate EM. The parameters are $\gamma_{nk} = 5$ and $a = 10$. The plot shows the probability π_1 of the first component versus iteration index. The starting value of π_1 was .3

A locally Lipschitz function is called *regular* if it admits a directional derivative at every point and if moreover this directional derivative coincides with Clarke's generalized directional derivative.

The Clarke subdifferential of f at θ is the convex set defined by

$$\partial f(\theta) = \{\eta \mid f^{\circ}(\theta, \omega) \geqslant \eta^t \omega, \ \forall \omega\}. \tag{6.22}$$

Proposition 6.1 *The function f is differentiable if and only if $\partial f(\theta)$ is a singleton.*

We now introduce another very important property of the Clarke subdifferential related to generalization of semicontinuity for set-valued maps.

Definition 6.2 *A set-valued map Φ is said to be outer-semicontinuous if its graph*

$$graph\ \Phi = \{(\theta, g) \mid g \in \Phi(\theta)\}$$

is closed, i.e. if for any sequence $(graph\Phi \ni)\ (\theta_n, g_n) \to (\theta^*, g^*)$ as $n \to +\infty$, then $(\theta^*, g^*) \in graph\Phi$.

One crucial property of the Clarke subdifferential is that it is outer-semicontinuous.

A point θ is said to be a *stationary point* of f if

$$0 \in \partial f(\theta).$$

Consider now the problem

$$\sup_{\theta \in \mathbb{R}^p} f(\theta)$$

subject to

$$g(\theta) = [g_1(\theta), \ldots, g_m(\theta)]^t \geqslant 0$$

where all the functions are locally Lipschitz from $\mathbb{R}^p$ to $\mathbb{R}$. Then, a necessary condition for optimality of θ is the Karush-Kuhn-Tucker condition, i.e. there exists a vector $u \in \mathbb{R}^m_+$ such that

$$0 \in \partial f(\theta) + \sum_{j=1}^{m} u_j \partial g_j(\theta).$$

Convex functions are in particular locally Lipschitz. The main references for these facts are Rockafellar (1970) and Hiriart-Urruty and Lemaréchal (1993).

Bibliography

[Akaike (1973)] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory (Tsahkadsor, 1971), pp. 267–281. Akadémiai Kiadó, Budapest.

[Alliney and Ruzinsky (1994)] Alliney, S. and Ruzinsky, S. A. (1994). An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation. IEEE Transactions on Signal Processing, 42 (3), 618–627.

[Barron (1999)] Barron, A. R. (1999). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. Bayesian statistics, 6 (Alcoceber, 1998), 27–52, Oxford University Press, New York, 1999.

[Berlinet and Roland (2007)] Berlinet, A. and Roland, Ch. (2007). Acceleration schemes with application to the EM algorithm. Computational Statistics and Data Analysis, 51, 3689-3702.

[Biernacki and Chrétien (2003)] Biernacki, C. and Chrétien, S. (2003). Degeneracy in the Maximum Likelihood Estimation of Univariate Gaussian Mixtures with EM. Statistics and Probability Letters, 61, 373-382.

[Candès and Plan (2009)] Candès, E. and Plan, Y. (2009). Near-ideal model selection by L1 minimization. The Annals of Statistics 37 (5), 2145–2177.

[Candès and Tao (2007)] Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. The Annals of Statistics 35 (6), 2313–2351.

[Celeux *et al.* (2001)] Celeux, G., Chrétien, S., Forbes, F. and Mkhadri, A. (2001). A Component-Wise EM Algorithm for Mixtures. Journal of Computational and Graphical Statistics, 10 (4), 697-712.

[Chrétien and Hero (2000)] Chrétien, S. and Hero, A. O. (2000). Kullback proximal algorithms for maximum-likelihood estimation. Information-theoretic imaging. IEEE Transactions on Information Theory 46 (5) 1800–1810.

[Chrétien and Hero (2008)] Chrétien, S. and Hero, A. O. (2008). On EM algorithms and their proximal generalizations. European Society for Applied and Industrial Mathematics Probability and Statistics 12, 308–326

[Clarke (1990)] Clarke, F. (1990). *Optimization and Nonsmooth Analysis*, Vol. 5, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, xii+308 pp. (p. 115).

[Cover and Thomas (1987)] Cover, T. and Thomas, J. (1987). *Elements of Information Theory*, Wiley, New York.

[Delyon, Lavielle and Moulines (1999)] Delyon, B., Lavielle, M. and Moulines, E.(1999). Convergence of a stochastic approximation version of the EM algorithm. The Annals of Statistics 27 (1), 94–128.

[Dempster, Laird, and Rubin (1977)] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39 (1), 1–38.

[Fan and Li (2001)] Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties, Journal of the American Statistical Association, 96, 1348–1360.

[Fan and Li (2002)] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. The Annals of Statistics, 30, 74–99.

[Fessler and Hero (1994)] Fessler, J. A. and Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. IEEE Transactions on Signal Processing, 42 (10), 2664–2677.

[Figueiredo and Nowak (2003)] Figueiredo, M. A. T. and Nowak, R. D. (2003). An EM algorithm for wavelet-based image restoration. IEEE Transactions on Image Processing, 12 (8), 906–916.

[Friedmand and Popescu (2003)] Friedmand, J. and Popescu, B. E. (2003). Importance Sampled Learning Ensembles, Journal of Machine Learning Research submitted.

[Green (1990)] Green, P. J. (1990). On the use of the EM algorithm for penalized likelihood estimation. J. Royal Statistical Society, Ser. B, 52 (2), pp. 443–452.

[Hero and Fessler (1995)] Hero, A. O. and Fessler, J. A. (1995). Convergence in norm for alternating expectation-maximization (EM) type algorithms. Statistica Sinica, 5 (1), 41–54.

[Hiriart-Urruty and Lemaréchal (1993)] Hiriart-Urruty, J. B. and Lemaréchal, C. (1993). Convex Analysis and Minimization Algorithms, Vol. 306 Grundlehren der mathematischen Wissenschaften, Springer

[Hunter and Lange (2004)] Hunter, D. R. and Lange, K. (2004). A Tutorial on MM Algorithms. The American Statistician, 58 (1), 30–37.

[Hunter and Li (2005)] Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. The Annals of Statistics, 33 (4), 1617–1642..

[Ibragimov and Has'minskii (1981)] Ibragimov, I. A.; Has'minskii, R. Z. (1981) Statistical estimation. Asymptotic theory. Translated from the Russian by Samuel Kotz. Applications of Mathematics, 16. Springer-Verlag, New York-Berlin. vii+403 pp.

[Johnstone and Silverman (2004)] Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. The Annals of Statistics, 32 (4), 1594–1649.

[Khalili and Chen (2007)] Khalili, A. and Chen, J. (2007). Variable Selection in Finite Mixture of Regression Models. Journal of the American Statistical Association, 102 (479), 1025-1038.

[Koh, Kim and Boyd (2007)] Koh, K., Kim, S.-J. and Boyd, S. (2007). An Interior-Point Method for Large-Scale l1-Regularized Logistic Regression. Journal of Machine Learning Research, 8, 1519–1555.

[Kuhn and Lavielle (2004)] Kuhn, E., and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. European Society for Applied and Industrial Mathematics Probability and Statistics 8, 115–131

[Lange (1995)] Lange, K. (1995). A quasi-newtonian acceleration of the EM algorithm. Statistica Sinica, 5 (1), 1–18.

[Liu, Rubin and Wu (1998)] Liu, C., Rubin, D. B. and Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. Biometrika 85 (4), 755–770.

[Martinet (1970)] Martinet, B. (1970). Régularisation d'inéquation variationnelles par approximations successives. Revue Francaise d'Informatique et de Recherche Operationnelle, 3, 154–179.

[McLachlan and Peel (2000)] McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Interscience, New York, 2000. xxii+419 pp.

[Minty (1962)] Minty, G. J. (1962). Monotone (nonlinear) operators in Hilbert space. Duke Mathematical Journal, 29, 341–346.

[Moreau (1965)] Moreau, J. J. (1965). Proximité et dualité dans un espace Hilbertien. Bulletin de la Société Mathématique de France, 93, pp. 273–299.

[Rockafellar (1970)] Rockafellar, R. T. (1970). Convex Analysis, Convex analysis. Princeton Mathematical Series, No. 28 Princeton University Press, Princeton, N.J. 1970 xviii+451 pp.

[Rockafellar (1976)] Rockafellar, R. T. (1976). Monotone operators and the proximal point algorithm. Society for Industrial and Applied Mathematics Journal on Control and Optimization, 14, 877–898.

[Rockafellar and Wets (2004)] Rockafellar, R. T. and Wets, R. J. B. (2004). Variational Analysis. Variational analysis. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 317. Springer-Verlag, Berlin, 1998. xiv+733 pp.

[Schwarz (1978)] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461–464.

[Tibshirani (1996)] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B, 58 (1), 267–288.

[Tipping (01)] Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 1 (3), 211–244.

[Varadhan and Roland (2007)] Varadhan, R. and Roland, Ch. (2007). Simple and Globally-Convergent Numerical Methods for Accelerating Any EM Algorithm. Scandinavian Journal of Statistics, 35 (2), 335–353.

[Wu (1983)] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. The Annals of Statistics, 11, 95–103.

Chapter E

Multivariate GARCH calibration via Bregman divergences

with Juan-Pablo Ortega

Abstract

This paper presents a global matrix formulation of the VEC model calibration problem and proposes a well adapted optimization method to solve it based on Bregman divergences. More especifically, the calibration method is articulated as an optimization problem with constraints that are formulated as matricial positive definiteness conditions which are sufficient requirements for the resulting model to be stationary and to exhibit well-defined conditional covariance matrices. The resulting optimization problem is solved by using local models that incorporate appropriate Bregman divergences that ensure that, at each iteration, the constraints are satisfied. Details on how to improve the performance of the method using quadratic BFGS corrections and trust-region algorithmics are provided, as well as a preliminary estimation technique that shows how to efficiently initialize the algorithm.

1 Introduction

Autoregressive conditionally heteroscedastic (ARCH) models [**?**] and their generalized counterparts (GARCH) [**?**] are standard econometric tools to capture the leptokurticity and the volatility clustering exhibited by financial time series. In the one dimensional situation, a large collection of models that account for various stylized features of financial returns is available, as well as model selection and calibration tools, and explicit characterizations of the conditions that ensure stationarity or the existence of higher moments. One of the advantages of GARCH models that makes them particularly useful is that once they have been calibrated they provide an estimate of the dynamical behavior of volatility which, in principle, is not directly observable.

The last remark makes desirable the extension of the GARCH prescription to the multivariate case since such a generalization provides a dynamical picture of the correlations between different assets which are of major importance for pricing, asset allocation, risk management, and hedging purposes.

This generalization is nevertheless not free from difficulties. The most general multivariate GARCH models are the VEC prescription proposed by Bollerslev et al [**?**] and the

BEKK model by Engle et al [**?**]; both families of models present satisfactory properties that match those found in univariate GARCH models, nevertheless their lack of parsimony, even in low dimensions makes them extremely difficult to calibrate; for example, VEC(1,1) models require $n(n+1)(n(n+1)+1)/2$ parameters, where n is the dimensionality of the modeling problem; BEKK(1,1,1) require $n(5n+1)/2$. Indeed, due to the high number of parameters needed, it is rare to find these models at work beyond two or three dimensions even when ad hoc calibration techniques are used; see for example [**?**].

The goal of the work that we present in this paper is increasing the range of dimensions in which VEC models can be calibrated in practice by improving the existing technology in two directions:

- Explicit matrix formulation of the model and of the associated stationarity and positivity constraints: the existing works in the literature usually proceed by expressing the constraints in terms of the entries of the parameter matrices [**?**]. A global matrix formulation is necessary in order to obtain a dimension independent encoding of the problem.

- Use of the optimization method developed by Nesterov in [**?**]. This method is of particular interest to us since it is a purely first order method. More explicitely, unlike the Newton algorithm or other more sophisticated line search techniques, Nesterov's method only requires the computation of the first derivative of the target function. This feature is of paramount importance in our situation as the high number of parameters makes unviable the use of Hessians. Additionally, if k denotes the iteration number in the optimization algorithm, the convergence rate of this method is proportional to $1/k^2$ instead of the $1/k$ rate provided by standard gradient descent techniques.

Notation and conventions: all along this paper, bold symbols like $\mathbf{r}$ denote column vectors, $\mathbf{r}^T$ denotes the transposed vector. Given a filtered probability space $(\Omega, \mathbb{P}, \mathcal{F}, \{\mathcal{F}_t\}_{t\in\mathbb{N}})$ and X, Y two random variables, we will denote by $E_t[X] := E[X|\mathcal{F}_t]$ the conditional expectation, $cov_n(X,Y) := cov(X,Y|\mathcal{F}_n) := E_t[XY] - E_t[X]E_t[Y]$ the conditional covariance, and by $var_t(X) := E_t[X^2] - E_t[X]^2$ the conditional variance. A discrete-time stochastic process $\{X_t\}_{t\in\mathbb{N}}$ is predictable when X_t is $\mathcal{F}_{t-1}$-measurable, for any $t \in \mathbb{N}$.

2 Preliminaries on matrices and matrix operators

Matrices: Let $n, m \in \mathbb{N}$ and denote by $\mathbb{M}_{n,m}$ the space of $n \times m$ matrices. When $n = m$ we will just write $\mathbb{M}_n$ to refer to the space of $n \times n$ square matrices. Unless specified otherwise, all the matrices in this paper will contain purely real entries. The equality $A = (a_{ij})$ denotes the matrix A with components $a_{ij} \in \mathbb{R}$. The symbol $\mathbb{S}_n$ denotes the subspace of $\mathbb{M}_n$ that contains all symmetric matrices

$$\mathbb{S}_n = \{A \in \mathbb{M}_m \mid A^T = A\}$$

and $\mathbb{S}_n^+$ (respectively $\mathbb{S}_n^-$) is the cone in $\mathbb{S}_n$ containing the positive (respectively negative) semidefinite matrices. The symbol $A \succeq 0$ (respectively $A \preceq 0$) means that A is positive (respectively negative) semidefinite.

We will consider $\mathbb{M}_{n,m}$ as an inner product space with the pairing

$$\langle A, B\rangle = trace(AB^T) \tag{2.1}$$

and denote by $\|A\| = \langle A, A\rangle^{\frac{1}{2}}$ the associated Frobenius norm. Given a linear operator $\mathcal{A} : \mathbb{M}_{n,m} \to \mathbb{M}_{p,q}$ we will denote by $\mathcal{A}^* : \mathbb{M}^*_{p,q} \to \mathbb{M}^*_{n,m}$ its adjoint with respect to the inner product (2.1).

The vec, vech, and math operators and their adjoints: The symbol $vec : \mathbb{M}_n \to \mathbb{R}^{n^2}$ denotes the operator that stacks all the columns of a matrix into a vector. Let $N = \frac{1}{2}n(n+1)$ and let $vech : \mathbb{S}_n \to \mathbb{R}^N$ be the operator that stacks only the lower triangular part, including the diagonal, of a symmetric matrix into a vector. The inverse of the vech operator will be denoted by $math : \mathbb{R}^N \to \mathbb{S}_n$.

Given $n \in \mathbb{N}$ and $N = \frac{1}{2}n(n+1)$, let $S = \{(i,j) \in \{1,\dots,n\}\times\{1,\dots,n\} \mid i \geqslant j\}$ we define $\sigma : S \to \{1,\dots,N\}$ as the map that yields the position of component $(i,j), i \geqslant j$, of any symmetric matrix in its equivalent vech representation. The symbol $\sigma^{-1} : \{1,\dots,N\} \to S$ will denote its inverse and $\widetilde{\sigma} : \{1,\dots,n\}\times\{1,\dots,n\} \to \{1,\dots,N\}$ the extension of σ defined by:

$$\widetilde{\sigma}(i,j) = \begin{cases} \sigma(i,j) & i \geqslant j \\ \sigma(j,i) & i < j. \end{cases} \tag{2.2}$$

The proof of the following result is provided in the Appendix.

Proposition 2.1 *Given $n \in \mathbb{N}$ and $N = \frac{1}{2}n(n+1)$, let $A \in \mathbb{S}_n$ and $m \in \mathbb{R}^N$ arbitrary. The following identities hold true:*

(i) $\langle vech(A), m\rangle = \frac{1}{2}\langle A + diag(A), math(m)\rangle$.

(ii) $\langle A, math(m)\rangle = 2\langle vech(A - \frac{1}{2}diag(A)), m\rangle$,

where $diag(A)$ denotes the diagonal matrix obtained out of the diagonal entries of A. Let $vech^* : \mathbb{R}^N \to \mathbb{S}_n$ and $math^* : \mathbb{S}_n \to \mathbb{R}^N$ be the adjoint maps of vech and math, respectively, then:

$$math^*(A) = 2\,vech\left(A - \frac{1}{2}diag(A)\right), \tag{2.3}$$

$$vech^*(m) = \frac{1}{2}\left(math(m) + diag(math(m))\right). \tag{2.4}$$

The operator norms of the mappings that we just introduced are given by:

$$\|vech\|_{op} = 1 \tag{2.5}$$

$$\|math\|_{op} = \sqrt{2} \tag{2.6}$$

$$\|vech^*\|_{op} = 1 \tag{2.7}$$

$$\|math^*\|_{op} = \sqrt{2} \tag{2.8}$$

$$\|diag\|_{op} = 1 \tag{2.9}$$

Block matrices and the Σ operator: let $n \in \mathbb{N}$ and $B \in \mathbb{M}_{n^2}$. The matrix B can be divided into n^2 blocks $B_{ij} \in \mathbb{M}_n$ and hence its components can be labeled using a blockwise notation by referring to the (k,l) element of the (i,j) block as $(B_{ij})_{kl}$. This notation makes particularly accessible the interpretation of B as the coordinate expression of a linear endomorphism of the tensor product space $\mathbb{R}^n \otimes \mathbb{R}^n$. Indeed if $\{\mathbf{e}_1,\dots,\mathbf{e}_n\}$ is the canonical basis or $\mathbb{R}^n$, we have

$$B(\mathbf{e}_i \otimes \mathbf{e}_k) = \sum_{j,l=1}^{n} (B_{ij})_{kl}(\mathbf{e}_j \otimes \mathbf{e}_l). \tag{2.10}$$

Definition 2.2 *Let $A \in \mathbb{M}_N$ with $N = \frac{1}{2}n(n+1)$. We define $\Sigma(A) \in \mathbb{S}_{n^2}$ blockwise using the expression*

$$\begin{cases} \text{If } k \geqslant l \quad (\Sigma(A)_{kl})_{ij} & = \begin{cases} \frac{1}{2}A_{\sigma(k,l),\sigma(i,j)}, & \text{if} \quad i > j \\ A_{\sigma(k,l),\sigma(i,j)}, & \text{if} \quad i = j \\ \frac{1}{2}A_{\sigma(k,l),\sigma(j,i)}, & \text{if} \quad i < j \end{cases} \\ \text{If } k \leqslant l \qquad \Sigma(A)_{kl} & = \Sigma(A)_{lk}, \end{cases} \tag{2.11}$$

*where σ is the map defined above that yields the position of component $(i,j), i \geqslant j$, of any symmetric matrix in its equivalent vech representation. By construction $(\Sigma(A)_{kl})_{ij}$ is symmetric with respect to transpositions in the (k,l) and (i,j) indices; this implies that $\Sigma(A)$ is both symmetric and blockwise symmetric. We will refer to any matrix in $\mathbb{S}_{n^2}$ with this property as n-**symmetric** and will denote the corresponding space by $\mathbb{S}^n_{n^2}$.*

The proofs of the next two results are provided in the Appendix.

Proposition 2.3 *Given $H \in \mathbb{S}_n$ and $A \in \mathbb{M}_N$, with $N = \frac{1}{2}n(n+1)$, the n-symmetric matrix $\Sigma(A) \in \mathbb{S}^n_{n^2}$ that we just defined satisfies:*

$$A\, vech(H) = vech(\Sigma(A) \bullet H), \tag{2.12}$$

where $\Sigma(A) \bullet H \in \mathbb{S}_n$ is the symmetric matrix given by

$$(\Sigma(A) \bullet H)_{kl} = \langle \Sigma(A)_{kl}, H \rangle = \mathrm{trace}(\Sigma(A)_{kl} H).$$

Proposition 2.4 *Let $\Sigma : \mathbb{M}_N \to \mathbb{M}_{n^2}$ be the operator defined in the previous proposition, $N = \frac{1}{2}n(n+1)$. Then, for any $\mathcal{B} \in M_{n^2}$, the corresponding dual map $\Sigma^* : \mathbb{M}_{n^2} \to \mathbb{M}_N$ is given by*

$$\Sigma^*(\mathcal{B}) = 2B - \widetilde{B}, \tag{2.13}$$

where $B, \widetilde{B} \in \mathbb{M}_N$ are the matrices defined by

$$B_{pq} = ((\mathbb{P}^n_{n^2}(\mathcal{B}))_{\sigma^{-1}(p)})_{\sigma^{-1}(q)}, \quad \text{and} \quad \widetilde{B}_{pq} = B_{pq}\delta_{pr_1(\sigma^{-1}(p)),pr_2(\sigma^{-1}(p))}.$$

The symbol $\mathbb{P}^n_{n^2}(\mathcal{B})$ denotes the orthogonal projection of $\mathcal{B} \in \mathbb{M}_{n^2}$ onto the space $\mathbb{S}^n_{n^2}$ of n-symmetric matrices that we spell out in Lemma 5.1. As we saw in Proposition 2.3, Σ maps into the space $\mathcal{S}^n_{n^2}$ of symmetric matrices; let $\widetilde{\Sigma} : \mathbb{M}_N \to \mathcal{S}^n_{n^2}$ be the map obtained out of Σ by restriction of its range. The map $\widetilde{\Sigma}$ is a bijection with inverse $\widetilde{\Sigma}^{-1} : \mathcal{S}^n_{n^2} \to \mathbb{M}_N$ given by

$$\left(\Sigma^{-1}(B)\right)_{p,q} = \left(B_{\sigma^{-1}(p)}\right)_{\sigma^{-1}(q)} \left(2 - \delta_{pr_1(\sigma^{-1}(q)),pr_2(\sigma^{-1}(q))}\right). \tag{2.14}$$

3 The VEC-GARCH model

Consider the n-dimensional conditionally heteroscedastic discrete-time process $\{\mathbf{z}_t\}$ determined by the relation

$$\mathbf{z}_t = H_t^{1/2}\boldsymbol{\varepsilon_t} \qquad \text{with} \qquad \{\boldsymbol{\varepsilon_t}\} \sim IIDN(\mathbf{0}, \boldsymbol{I}_n).$$

In this expression, $\{H_t\}$ denotes a predictable matrix process, that is for each $t \in \mathbb{N}$, the matrix random variable H_t is $\mathcal{F}_{t-1}$-measurable, and $H_t^{1/2}$ is a square root of H_t, hence it

satisfies $H_t^{1/2}(H_t^{1/2})^T = H_t$. In these conditions it is easy to show that the conditional mean $E_t[\mathbf{z}_t] = \mathbf{0}$ and that the conditional covariance matrix process of $\{\mathbf{z}_t\}$ coincides with $\{H_t\}$.

Different prescriptions for the time evolution of the conditional covariance matrix $\{H_t\}$ determine different vector conditional heteroscedastic models. In this paper we will focus on the **VEC-GARCH model** (just VEC in what follows). This model was introduced in [**?**] as the direct generalization of the univariate GARCH model [**?**] in the sense that every conditional variance and covariance is a function of all lagged conditional variances and covariances as well as all squares and cross-products of the lagged time series values. More specifically, the VEC(q,p) model is determined by

$$\boldsymbol{h}_t = \mathbf{c} + \sum_{i=1}^{q} A_i \boldsymbol{\eta}_{t-i} + \sum_{i=1}^{p} B_i \boldsymbol{h}_{t-i},$$

where $\boldsymbol{h}_t := vech(H_t)$, $\boldsymbol{\eta}_t := \mathbf{z}_t\mathbf{z}^T$, $\mathbf{c}$ is a N-dimensional vector, with $N := n(n+1)/2$ and $A_i, B_i \in \mathbb{M}_N$.

In the rest of the paper we will restrict to the case $p = q = 1$, that is:

$$\left\{ \begin{array}{rcl} \mathbf{z}_t & = & H_t^{1/2}\boldsymbol{\varepsilon_t} \qquad \text{with} \qquad \{\boldsymbol{\varepsilon_t}\} \sim IIDN(\mathbf{0}, \boldsymbol{I}_n), \\ \boldsymbol{h}_t & = & \mathbf{c} + A\boldsymbol{\eta}_{t-1} + B\boldsymbol{h}_{t-1}. \end{array} \right. \tag{3.15}$$

In this case the model needs $N(2N+1) = \frac{1}{2}(n^2+n)(n^2+n+1)$ parameters for a complete specification.

Positivity and stationarity constraints

The general prescription for the VEC model spelled out in (3.15) does not guarantee that it has stationary solutions. Moreover, as we saw above, the resulting matrices $\{H_t\}_{t\in\mathbb{N}}$ are the conditional covariance matrices of the resulting process and therefore, additional constraints should be imposed on the parameter matrices $\mathbf{c}, A$, and B in order to ensure that they are symmetric and positive semidefinite. Unlike the situation encountered in the one-dimensional case, necessary and sufficient conditions for positivity and stationarity are very difficult to find and we will content ourselves with sufficient specifications.

Positivity constraints: we will use the sufficient conditions introduced by Gourieroux in [**?**] that, as we show in the next proposition, can be explicitly formulated using the map Σ introduced in Definition 2.2.

Proposition 3.1 *If the parameter matrices $\boldsymbol{c}, A$, and B in (3.15) are such that $math(\boldsymbol{c}), \Sigma(A)$, and $\Sigma(B)$ are positive semidefinite then so are the resulting conditional covariance matrices $\{H_t\}_{t\in\mathbb{N}}$, provided the initial condition H_0 is positive semidefinite.*

Second order stationarity constraints: Gourieroux [**?**] has stated sufficient conditions in terms of the spectral radius of $A+B$ that we will make more restrictive in order to ensure the availability of a formulation in terms of positive semidefiniteness constraints.

Proposition 3.2 *The VEC model specified in (3.15) admits a unique second order stationary solution if all the eigenvalues of $A+B$ lie strictly inside the unit circle. This is always the case whenever the top singular eigenvalue $\sigma_{max}(A+B)$ of $A+B$ is smaller than one or, equivalently, when the matrix $\mathbb{I}_N - (A+B)(A+B)^T$ is positive definite. If any of these conditions is satisfied, the marginal variance of the model is given by*

$$\Gamma(0) = math(E[\boldsymbol{h}_t]) = math((\mathbb{I}_N - A - B)^{-1}\boldsymbol{c}). \tag{3.16}$$

The likelihood function, its gradient, and computability constraints

Given a sample $\mathbf{z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_T\}$, the quasi-loglikelihood associated to (3.15) is:

$$logL(\mathbf{z}; \boldsymbol{\theta}) = -\frac{TN}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{T} \log(\det H_t) - \frac{1}{2} \sum_{t=1}^{T} \mathbf{z}_t^T H_t^{-1} \mathbf{z}_t \tag{3.17}$$

where $\boldsymbol{\theta} := (\mathbf{c}, A, B)$. In this expression, the matrices H_t are constructed out of $\boldsymbol{\theta}$ and the sample $\mathbf{z}$ using the second expression in (3.15). This implies that the dependence of $logL$ on $\boldsymbol{\theta}$ takes place through the matrices H_t. Notice that these matrices are well defined once initial values H_0 and $\mathbf{z}_0$ have been fixed. This initial values are usually taken out of a presample; if this is not available it is customary to take the mean values associated to the stationary model, namely $\mathbf{z} = \mathbf{0}$ and $H_0 = math((\mathbb{I}_N - A - B)^{-1}\mathbf{c})$ (see (3.16)). Once the initial conditions have been fixed, it can be shown by induction that

$$\mathbf{h}_t = \left(\sum_{i=0}^{t-1} B^i \right) \mathbf{c} + \sum_{i=0}^{t-1} B^i A \boldsymbol{\eta}_{t-i-1} + B^t \mathbf{h}_0. \tag{3.18}$$

The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is the value that maximizes (3.17) for a given sample $\mathbf{z}$. The search of that extremal is carried out using an optimization algorithm that we will discuss later on in the paper and that requires the gradient $\nabla_{\boldsymbol{\theta}} logL(\mathbf{z}; \boldsymbol{\theta})$ of $logL$. In order to compute it we write the total quasi-loglikelihood as a sum of T conditional loglikelihoods

$$l_t(\mathbf{z}_t; A, B, c) = -\frac{N}{2} \log 2\pi - \frac{1}{2} \log(\det H_t) - \frac{1}{2} \mathbf{z}_t^T H_t^{-1} \mathbf{z}_t$$

A lengthy calculation shows that:

$$\nabla_{\mathbf{c}} l_t = \left[(\gamma_t - \Gamma_t)^T \sum_{i=0}^{t-1} B^i \right]^T, \tag{3.19}$$

$$\nabla_A l_t = \left[\sum_{i=0}^{t-1} \boldsymbol{\eta}_{t-i-1} (\gamma_t - \Gamma_t)^T B^i \right]^T, \tag{3.20}$$

$$\nabla_B l_t = \left[\sum_{i=0}^{t-1} \left[\sum_{j=0}^{i-1} B^j (\mathbf{c} + A\boldsymbol{\eta}_{t-i-1}) (\gamma_t - \Gamma_t)^T B^{i-j-1} + B^j \mathbf{h}_0 (\gamma_t - \Gamma_t)^T B^{t-j-1} \right] \right]^T \tag{3.21}$$

where

$$\Gamma_t := \frac{1}{2} math^*(H_t^{-1}), \quad \gamma_t := \frac{1}{2} math^*(\Lambda_t), \quad \text{and} \quad \Lambda_t := H_t^{-1} \mathbf{z}_t \mathbf{z}_t^T H_t^{-1}.$$

These formulas for the gradient were obtained by using the explicit expression of the conditional covariance matrices (3.18) in terms of the sample elements and the coefficient matrices. Such a closed form expression is not always available as soon as the model becomes slightly more complicated; for example, if one adds to the model (3.15) a drift term like in [**?**] for the one dimensional GARCH case, an expression like (3.18) ceases to exist. That is why, in the next proposition, we introduce an alternative iterative method that can be extended to more general models, it is well adapted to its use under the form of a computer code and, more importantly, suggests the introduction of an additional calibration constraint that noticeably shortens the computation time needed for its numerical evaluation.

Proposition 3.3 *Let $\boldsymbol{z} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_T\}$ be a sample, $\boldsymbol{\theta} := (\boldsymbol{c}, A, B)$, and let $logL(\boldsymbol{z}; \boldsymbol{\theta})$ be the quasi-loglikelihood introduced in (3.17). Then, for any component θ of the three-tuple $\boldsymbol{\theta}$, we have*

$$\nabla_{\boldsymbol{\theta}} \log L = \sum_{t=1}^{T} \nabla_{\boldsymbol{\theta}} l_t = \sum_{t=1}^{T} T^*_{\boldsymbol{\theta}} H_t \cdot \nabla_{H_t} l_t, \quad \textit{where} \tag{3.22}$$

$$\nabla_{H_t} l_t = -\frac{1}{2}\left[H_t^{-1} - H_t^{-1} \boldsymbol{z}_t \boldsymbol{z}_t^T H_t^{-1}\right], \tag{3.23}$$

and the differential operators $T^*_\theta H_t$ are determined by the recursions:

$$\begin{aligned}
T^*_{\boldsymbol{c}} H_t \cdot \Delta &= math^*(\Delta) + T^*_{\boldsymbol{c}} H_{t-1} \cdot vech^*(B^T math^*(\Delta)), && (3.24)\\
T^*_A H_t \cdot \Delta &= math^*(\Delta) \cdot \boldsymbol{\eta}_{t-1}^T + T^*_A H_{t-1} \cdot vech^*(B^T math^*(\Delta)), && (3.25)\\
T^*_B H_t \cdot \Delta &= math^*(\Delta) \cdot vech(H_{t-1})^T + T^*_B H_{t-1} \cdot vech^*(B^T math^*(\Delta)), && (3.26)
\end{aligned}$$

with $\boldsymbol{\eta}_t = \boldsymbol{z}_t \boldsymbol{z}^T$, $\Delta \in \mathbb{S}_n$ and setting $T^*_{\boldsymbol{c}} H_0 = \boldsymbol{0}$, $T^*_A H_0 = T^*_B H_0 = \boldsymbol{0}$. The operators $T^*_\theta H_t$ constructed in (3.24)–(3.26) are the adjoints of the partial tangent maps $T_{\boldsymbol{c}} H_t : \mathbb{R}^N \to \mathbb{S}_n$, $T_A H_t : M_N \to \mathbb{S}_n$, and $T_B H_t : M_N \to \mathbb{S}_n$ to $H_t(\boldsymbol{c}, A, B) := math(\boldsymbol{h}_t(\boldsymbol{c}, A, B))$, with $\boldsymbol{h}_t(\boldsymbol{c}, A, B)$ as defined in (3.18).

Whenever we deal with a long time series sample, the computation of the gradient (3.22) may turn out numerically very expensive since it consists of the sum of T terms $T^*_\theta H_t \cdot \nabla_{H_t} l_t$, each of which is made of the sum of the t terms recursively defined in (3.24)–(3.26). A major simplification can be obtained if we restrict ourselves in the calibration process to matrices B whose top eigenvalue is in norm smaller than one. The defining expressions for the differential operators $T^*_\theta H_t$ show that in that situation, only a certain number of iterations, potentially small, is needed to compute the gradients with a prescribed precision. This is particularly visible in the expressions (3.19)–(3.21) where the dependence on the powers of B makes very small many of the involved summands whenever the spectrum of B is strictly contained in the unit disk. This is the reason why we will impose this as an additional calibration constraint. The details of this statement are spelled out in the proposition below that we present after the summary of the constraints that we will impose all along the paper on the model (3.15):

(SC) Stationarity constraints: $\mathbb{I}_N(1 - \varepsilon_{AB}) - (A + B)(A + B)^T \succeq 0$ for some small $\varepsilon_{AB} > 0$.

(PC) Positivity constraints: $math(\mathbf{c}) - \varepsilon_{\mathbf{c}} \mathbb{I}_n \succeq 0$, $\Sigma(A) - \varepsilon_A \mathbb{I}_{n^2} \succeq 0$, and $\Sigma(B) - \varepsilon_B \mathbb{I}_{n^2} \succeq 0$, for some small $\varepsilon_A, \varepsilon_B, \varepsilon_{\mathbf{c}} > 0$.

(CC) Computability constraints: $\mathbb{I}_N(1 - \widetilde{\varepsilon}_B) - BB^T \succeq 0$ for some small $\widetilde{\varepsilon}_B > 0$.

Proposition 3.4 *Let $t \in \mathbb{N}$ be a fixed lag and let $T^*_\theta H_t$ be the differential operators defined by applying t times the recursions (3.24)-(3.26). Consider now the operators $T^*_\theta H^k_t$ obtained by truncating the recursions (3.24)-(3.26) after k iterations, $k < t$. If we assume that the coefficients $\boldsymbol{c}, A$, and B satisfy the constraints **(SC)**, **(PC)**, and **(CC)** then the error committed in the truncations can be estimated using the following inequalities satisfied by*

the operator norms:

$$\|T_{\boldsymbol{c}}^* H_t - T_{\boldsymbol{c}}^* H_t^k\|_{op} \leqslant \frac{2(1-\widetilde{\varepsilon}_B)^k}{\widetilde{\varepsilon}_B}, \tag{3.27}$$

$$\|E\left[T_A^* H_t - T_A^* H_t^k\right]\|_{op} \leqslant \frac{2(1-\widetilde{\varepsilon}_B)^k \|\boldsymbol{c}\|}{\varepsilon_{AB}}, \tag{3.28}$$

$$\|E\left[T_B^* H_t - T_B^* H_t^k\right]\|_{op} \leqslant \frac{2(1-\widetilde{\varepsilon}_B)^k \|\boldsymbol{c}\|}{\varepsilon_{AB}}. \tag{3.29}$$

Notice that the last two inequalities estimate the error committed in mean. As consequence of these relations, if we allow a maximum expected error δ in the computation of the gradient (3.22) then a lower bound for the number k of iterations that need to be carried out in (3.24)–(3.26) is:

$$k = \max\left\{\frac{\log\left(\frac{\widetilde{\varepsilon}_B \delta}{2}\right)}{\log(1-\widetilde{\varepsilon}_B)}, \frac{\log\left(\frac{\widetilde{\varepsilon}_B \varepsilon_{AB} \delta}{2\varepsilon_c}\right)}{\log(1-\widetilde{\varepsilon}_B)}\right\}. \tag{3.30}$$

Remark 3.5 The estimate (3.30) for the minimum number of iterations needed to reach a certain precision in the computation of the gradient is by no means sharp. Numerical experiments show that the figure produced by this formula is in general too conservative. Nevertheless, this expression is still very valuable for it explicitly shows the pertinence of the computability constraint **(CC)**.

Remark 3.6 We emphasize that the constraints **(SC)**, **(PC)**, and **(CC)** are sufficient conditions for stationarity, positivity, and computability, respectively, but by no means necessary. For example **(SC)** and **(CC)** could be replaced by the more economical (but also more restrictive) condition that imposes $A, B \in \mathbb{S}_N^+$ with $\lambda_{max}(A+B) \leqslant (1-\varepsilon_{AB})$. In this situation it can be easily shown that $\lambda_{max}(B) < 1$ and hence the computability constrained is automatically satisfied.

4 Calibration via Bregman matrix divergences

In this section we present an efficient optimization method that, given a sample $\mathbf{z}$, provides the parameter value $\widehat{\boldsymbol{\theta}}$ corresponding to the VEC(1,1) model that fits it best by maximizing the quasi-loglikelihood (3.17) subjected to the constraints **(SC)**, **(PC)**, and **(CC)**. It can be proved under certain regularity hypotheses (see [**?**, page 119]) that the quasi-loglikelihood estimator $\widehat{\boldsymbol{\theta}}$ is consistent and asymptotically normal:

$$\sqrt{T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow[dist]{} N(0, \Omega_0) \quad \text{where} \quad \Omega_0 = A_0^{-1} B_0 A_0^{-1}, \quad \text{with}$$

$$A_0 = E_{\boldsymbol{\theta}_0}\left[-\frac{\partial^2 l_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right] \quad \text{and} \quad B_0 = E_{\boldsymbol{\theta}_0}\left[\frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\frac{\partial l_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^T}\right].$$

These matrices are usually consistently estimated by replacing the expectations by their empirical means and the true value of the parameter $\boldsymbol{\theta}_0$ by the estimator $\widehat{\boldsymbol{\theta}}$:

$$\widehat{A}_0 = -\frac{1}{T}\sum_{i=1}^{T}\frac{\partial^2 l_t(\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}, \qquad \widehat{B}_0 = \frac{1}{T}\sum_{i=1}^{T}\frac{\partial l_t(\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\frac{\partial l_t(\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}^T}.$$

Constrained optimization via Bregman divergences

The optimization method that we will be carrying out to maximize the quasi-loglikelihood is based on the use of **Burg's matrix divergence**. This divergence is presented, for example, in [**?**] and it is a particular instance of a Bregman divergence. Bregman divergences are of much use in the context of machine learning (see for instance [**?**, **?**] and references therein). In our situation we have opted for this technique as it allows for a particularly efficient treatment of the constraints in our problem, avoiding the need to solve additional secondary optimization problems that appear, for example, had we used Lagrange duality; even though the constraints that we handle admit a simple and explicit conic formulation well adapted to the use of Lagrange multipliers, the associated dual optimization problem is this case of difficulty comparable to that of the primal so avoiding this extra step is a major advantage.

Definition 4.1 *Let $X, Y \in \mathbb{S}_n$ and $\phi : \mathbb{S}_n \longrightarrow \mathbb{R}$ a strictly convex differentiable function. The **Bregman matrix divergence** associated to ϕ is defined by*

$$D_\phi(X, Y) := \phi(X) - \phi(Y) - \operatorname{trace}\left(\nabla\phi(Y)^T(X - Y)\right).$$

Bregman divergences are used to measure distance between matrices. Indeed, if we take the squared Frobenius norm as the function ϕ, that is $\phi(X) := \|X\|^2$, then $D_\phi(X, Y) := \|X - Y\|^2$. Other example is the **von Neumann divergence** which is the Bregman divergence associated to the entropy of the eigenvalues of a positive definite matrix; more explicitly, if X is a positive definite matrix with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, then $\phi(X) := \sum_{i=1}^n (\lambda_i \log \lambda_i - \lambda_i)$. In our optimization problem we will be using **Burg's matrix divergence** (also called the **LogDet divergence** or **Stein's loss** in the statistics literature [**?**]) which is the Bregman divergence obtained out of the Burg entropy of the eigenvalues of a positive definite matrix, that is $\phi(X) := -\sum_{i=1}^n \log \lambda_i$, or equivalently $\phi(X) := -\log\det(X)$. The resulting Bregman divergence over positive definite matrices is

$$D_B(X, Y) := \operatorname{trace}(XY^{-1}) - \log\det(XY^{-1}) - n. \tag{4.31}$$

The three divergences that we just introduced are examples of **spectral** divergences, that is, the function ϕ that defines them can be written down as the composition $\phi = \varphi \circ \lambda$, where $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}$ is differentiable strictly convex function and $\lambda : \mathbb{S}_n \longrightarrow \mathbb{R}^n$ is the function that lists the eigenvalues of X in algebraically decreasing order. It can be seen (see Appendix A in [**?**]) that spectral Bregman matrix divergences are invariant by orthogonal conjugations, that is, for any orthogonal matrix $Q \in \mathbb{O}_n$:

$$D_\phi(Q^T X Q, Q^T Y Q) = D_\phi(X, Y).$$

Burg divergences are invariant by an even larger group since

$$D_B(M^T X M, M^T Y M) = D_B(X, Y),$$

for any square non-singular matrix M. Additionally, for any non-zero scalar α:

$$D_B(\alpha X, \alpha Y) = D_B(X, Y).$$

The use of Bregman divergences in matrix constrained optimization problems is substantiated by replacing the quadratic term in the local model, that generally uses the Frobenius distance, by a Bregman divergence that places the set outside the constraints at an infinite distance. More explicitly, suppose that the constraints of a optimization problem are formulated as a positive definiteness condition $A \succeq 0$ and that we want to find

$$\underset{A \succeq 0}{arg\,min}\ f(A),$$

by iteratively solving the optimization problems associated to penalized local models of the form

$$f_{A^{(n)}}(A) := f\left(A^{(n)}\right) + \left\langle \nabla f\left(A^{(n)}\right), A - A^{(n)} \right\rangle + \frac{L}{2} D_\phi(A, A^{(n)}). \tag{4.32}$$

If in this local model we take $\phi(X) := \|X\|^2$ and the elastic penalization constant L is small enough, the minimum $arg\,min_{A \succeq 0}\, f_{A^{(n)}}(A)$ is likely to take place outside the constraints. However, if we use Burg's divergence D_B instead, and $A^{(n)}$ is positive definite, then so is $arg\,min_{A \succeq 0}\, f_{A^{(n)}}(A)$ for no matter what value of the parameter L. This is so because as A approaches the constraints, the term $D_\phi(A, A^{(n)})$ becomes increasingly close to infinity producing the effect that we just described. The end result of using Bregman divergences is that they reduce a constrained optimization problem to a series of local unconstrained ones.

The local approximation for VEC models

Before we tackle the VEC calibration problem, we add to **(SC)**, **(PC)**, and **(CC)** a fourth constraint on the variable $\mathbf{c} \in \mathbb{R}^N$ that makes compact the optimization domain:

(KC) Compactness constraint: $K\mathbb{I}_N - math(\mathbf{c}) \succeq 0$ for some $K \in \mathbb{R}$.

In practice the constant K is taken as a multiple of the Frobenius norm of the covariance matrix of the sample. This is a reasonable choice since by (3.16), in the stationary regime $\mathbf{c} = (\mathbb{I}_N - A - B)vech(\Gamma(0))$; moreover, by the constraint **(SC)** and (2.5) we have

$$\|\mathbf{c}\| = \|(\mathbb{I}_N - A - B)vech(\Gamma(0))\| \leqslant \|\mathbb{I}_N - A - B\|_{op}\|vec\|_{op}\|\Gamma(0)\| \leqslant 2\|\Gamma(0)\|.$$

Now, given a sample $\mathbf{z}$ and a starting value for the parameters $\boldsymbol{\theta}_0 = (\mathbf{c}_0, A_0, B_0)$, our goal is finding the minimum of minus the quasi-loglikelihood $f(\boldsymbol{\theta}) := -logL(\mathbf{z}; \boldsymbol{\theta})$, subjected to the constraints **(SC)**, **(PC)**, **(CC)**, and **(KC)**. We will worry about the problem of finding a preliminary estimation $\boldsymbol{\theta}_0$ later on in Section 4. As we said before, our method is based on recursively optimizing penalized local models that incorporate Bregman divergences that ensure that the constraints are satisfied. More especifically, the estimate of the optimum $\boldsymbol{\theta}^{(n+1)}$ after n iterations is obtained by solving

$$\boldsymbol{\theta}^{(n+1)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^N \times \mathbb{M}_N \times \mathbb{M}_N,}{arg\,min} \tilde{f}^{(n)}(\boldsymbol{\theta}), \tag{4.33}$$

where $\tilde{f}^{(n)}$ is defined by:

$$\begin{aligned}
\tilde{f}^{(n)}(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}^{(n)}) + \langle \nabla f(\boldsymbol{\theta}^{(n)}), \boldsymbol{\theta} - \boldsymbol{\theta}^{(n)} \rangle + \frac{L_1}{2} D_B(\mathbb{I}_N - (A+B)^T(A+B), \mathbb{I}_N - (A^{(n)} + B^{(n)})^T(A^{(n)} + B^{(n)})) \\
&+ \frac{L_2}{2} D_B(\Sigma(A), \Sigma(A^{(n)})) + \frac{L_3}{2} D_B(\Sigma(B), \Sigma(B^{(n)})) + \frac{L_4}{2} D_B(\mathbb{I}_N - B^T B, \mathbb{I}_N - B^{(n)\,T} B^{(n)}) \\
&+ \frac{L_5}{2} D_B(math(\mathbf{c}), math(\mathbf{c}^{(n)})) + \frac{L_6}{2} D_B(K\mathbb{I}_N - math(\mathbf{c}), K\mathbb{I}_N - math(\mathbf{c}^{(n)})).
\end{aligned} \tag{4.34}$$

Notice that for the sake of simplicity we have incorporated the constraints in the divergences with the constraint tolerances $\varepsilon_{AB}, \varepsilon_A, \varepsilon_B, \widetilde{\varepsilon}_B$, and $\varepsilon_{\mathbf{c}}$ set equal to zero.

The local optimization problem in (4.33) is solved by finding the value $\boldsymbol{\theta}_0$ for which

$$\nabla \tilde{f}^{(n)}(\boldsymbol{\theta}_0) = 0. \tag{4.35}$$

A long but straightforward computation shows that the gradient $\nabla \tilde{f}^{(n)}(\boldsymbol{\theta})$ is given by the expressions:

$$\begin{aligned}
\nabla_A \tilde{f}^{(n)}(\boldsymbol{\theta}) &= \nabla_A f(\boldsymbol{\theta}^{(n)}) - L_1(A+B)\left(\left(\mathbb{I}_N - (A^{(n)}+B^{(n)})^T(A^{(n)}+B^{(n)})\right)^{-1} - \left(\mathbb{I}_N - (A+B)^T(\right.\\
&\quad + \frac{L_2}{2}\Sigma^*\left(\Sigma(A^{(n)})^{-1} - \Sigma(A)^{-1}\right),\\
\nabla_B \tilde{f}^{(n)}(\boldsymbol{\theta}) &= \nabla_B f(\boldsymbol{\theta}^{(n)}) - L_1(A+B)\left(\left(\mathbb{I}_N - (A^{(n)}+B^{(n)})^T(A^{(n)}+B^{(n)})\right)^{-1} - \left(\mathbb{I}_N - (A+B)^T(\right.\\
&\quad + \frac{L_3}{2}\Sigma^*\left(\Sigma(B^{(n)})^{-1} - \Sigma(B)^{-1}\right) - L_4 B\left(\left(\mathbb{I}_N - B^{(n)\,T}B^{(n)}\right)^{-1} - \left(\mathbb{I}_N - B^T B\right)^{-1}\right),\\
\nabla_{\mathbf{c}} \tilde{f}^{(n)}(\boldsymbol{\theta}) &= \nabla_{\mathbf{c}} f(\boldsymbol{\theta}^{(n)}) + \frac{L_5}{2} math^*\left(math(\mathbf{c}^{(n)})^{-1} - math(\mathbf{c})^{-1}\right)\\
&\quad - \frac{L_6}{2} math^*\left((K\mathbb{I}_N - math(\mathbf{c}^{(n)}))^{-1} - (K\mathbb{I}_N - math(\mathbf{c}))^{-1}\right),
\end{aligned}$$

where $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^{(n)}) = -\nabla_{\boldsymbol{\theta}} logL(\mathbf{z};\boldsymbol{\theta}^{(n)})$ is provided by the expressions in Proposition 3.3.

Solving the local approximation problem by space augmentation

In order to solve the local approximation problem, we will solve a intermediate system defined on an augmented space. Using an augmented space will be important from a numerical viewpoint. Indeed, using Bregman divergences implicitely enforces the semi-definiteness of the involved matrices. However, solving the local approximation problem can only be performed using iterative schemes for which the positive semi-definiteness constraints are difficult to preserve along the iterations. Choosing an augmented space where the semi-definiteness constraint can be easily preserved may improve the algorithm's behavior drastically. In the present study, we chose the following augmented system of equations:

$$g^{(n)}(\vartheta) := (g_j^{(n)}(\vartheta))_{j=1,\ldots,7} = 0, \tag{4.39}$$

with

$$\begin{aligned}
g_1^{(n)}(\vartheta) &= g_1^{(n)} - L_1 D\left(\left(\mathbb{I}_N - D^{(n)\,T}D^{(n)}\right)^{-1} - \left(\mathbb{I}_N - D^T D\right)^{-1}\right)\\
&\quad + \frac{L_2}{2}\Sigma^*\left((E^{(n)})^{-1} - (E)^{-1}\right), &(4.40)\\
g_2^{(n)}(\vartheta) &= g_2^{(n)} - L_1 D\left(\left(\mathbb{I}_N - D^{(n)\,T}D^{(n)}\right)^{-1} - \left(\mathbb{I}_N - D^T D\right)^{-1}\right)\\
&\quad + \frac{L_3}{2}\Sigma^*\left((F^{(n)})^{-1} - (F)^{-1}\right) - L_4 B\left(\left(\mathbb{I}_N - B^{(n)\,T}B^{(n)}\right)^{-1} - \left(\mathbb{I}_N - B^T B\right)^{-1}\right), &(4.41)\\
g_3^{(n)}(\vartheta) &= g_3^{(n)} + \frac{L_5}{2} math^*\left(math(\mathbf{c}^{(n)})^{-1} - math(\mathbf{c})^{-1}\right)\\
&\quad - \frac{L_6}{2} math^*\left((K\mathbb{I}_N - math(\mathbf{c}^{(n)}))^{-1} - (K\mathbb{I}_N - math(\mathbf{c}))^{-1}\right), &(4.42)\\
g_4^{(n)}(\vartheta) &= A + B - D &(4.43)\\
g_5^{(n)}(\vartheta) &= \Sigma(A) - E &(4.44)\\
g_6^{(n)}(\vartheta) &= \Sigma(B) - F &(4.45)\\
g_7^{(n)}(\vartheta) &= G - math(\mathbf{c}), &(4.46)
\end{aligned}$$

$$g_1^{(n)} = \nabla_A f(\boldsymbol{\theta}^{(n)}) \quad (4.47)$$
$$g_2^{(n)} = \nabla_B f(\boldsymbol{\theta}^{(n)}) \quad (4.48)$$
$$g_3^{(n)} = \nabla_c f(\boldsymbol{\theta}^{(n)}) \quad (4.49)$$
$$(4.50)$$

and $\vartheta = (A, B, c, D, E, F, G)$. The differentials of $g_j^{(n)}(\vartheta)$, $j = 1, \ldots, 7$ are given by

$$\begin{aligned} T_{g_1^{(n)}}(\vartheta) \cdot H &= -L_1 H_D \left(\left(\mathbb{I}_N - D^{(n)^T} D^{(n)} \right)^{-1} - \left(\mathbb{I}_N - D^T D \right)^{-1} \right) \\ &\quad + L_1 D \left(\left(\mathbb{I}_N - D^T D \right)^{-1} \left(H_D^T D + D^T H_D \right) \left(\mathbb{I}_N - D^T D \right)^{-1} \right) \\ &\quad + \frac{L_2}{2} \Sigma^* \left(E^{-1} H_E E^{-1} \right), \quad (4.51) \\ T_{g_2^{(n)}}(\vartheta) \cdot H &= -L_1 H_D \left(\left(\mathbb{I}_N - D^{(n)^T} D^{(n)} \right)^{-1} - \left(\mathbb{I}_N - D^T D \right)^{-1} \right) \\ &\quad + \frac{L_3}{2} \Sigma^* \left(F^{-1} H_F F^{-1} \right) \quad (4.52) \\ &\quad + L_4 B \left(\left(\mathbb{I}_N - B^T B \right)^{-1} \left(H_B^T B + B^T H_B \right) \left(\mathbb{I}_N - B^T B \right)^{-1} \right), \quad (4.53) \\ T_{g_3^{(n)}}(\vartheta) \cdot H &= \frac{L_5}{2} math^* \left(math(\mathbf{c})^{-1} \; math(h_{\mathbf{c}}) \; math(\mathbf{c})^{-1} \right) \\ &\quad \frac{L_6}{2} math^* \left((K\mathbb{I}_N - math(\mathbf{c}))^{-1} \; math(h_{\mathbf{c}}) \; (K\mathbb{I}_N - math(\mathbf{c}))^{-1} \right) \quad (4.54) \\ T_{g_4^{(n)}}(\vartheta) \cdot H &= H_A + H_B - H_D \quad (4.55) \\ T_{g_5^{(n)}}(\vartheta) \cdot H &= \Sigma(H_A) - H_E \quad (4.56) \\ T_{g_6^{(n)}}(\vartheta) \cdot H &= \Sigma(H_B) - H_F \quad (4.57) \\ T_{g_7^{(n)}}(\vartheta) \cdot H &= H_G - math(h_{\mathbf{c}}), \quad (4.58) \end{aligned}$$

where $H = (H_A, H_B, h_{\mathbf{c}}, H_D, H_E, H_F, H_G)$. The simplest proposal for solving the augmented system (4.39) is to use Newton's method. Newton's method can be written as

$$\vartheta^{(l+1)} = \vartheta^{(l)} - H^{(l)} \quad (4.59)$$

with initial value $\vartheta^{(0)} = (A^{(n)}, B^{(n)}, c^{(n)}, D^{(n)}, E^{(n)}, F^{(n)}, G^{(n)})$, where $H^{(l)}$ is the solution of

$$T_{g^{(n)}}(\vartheta^{(l)}) \cdot H = g^{(n)}, \quad (4.60)$$

where $g^{(n)} = (g_1^{(n)}, g_2^{(n)}, g_3^{(n)}, 0, 0, 0, 0)$. A more appropriate method might be a projected vection of Newton's method, where (4.61) is replaced with

$$\vartheta^{(l+1)} = P_+ \left(\vartheta^{(l)} - H^{(l)} \right) \quad (4.61)$$

where P_+ is a projection operator defined by

$$P_+(A, B, \mathbf{c}, D, E, F, G) = (A, B, \mathbf{c}, P_{\succeq}(D), P_{\succeq}(E), P_{\succeq}(F), P_{\succeq}(G)) \quad (4.62)$$

and where P_{succeq} is the projection onto the cone of semidefinite matrices (with space dimensions implicitely defined by the matrix onto which it applies).

Performance improvement: BFGS and trust-region corrections

The speed of convergence of the calibration algorithm presented in the previous section can be significantly increased by enriching the local model with a quadratic BFGS (Broyden-Fletcher-Goldfarb-Shanno) type term and by only accepting steps of a certain quality measured by the ratio between the actual descent and that predicted by the local model (see [**?**] and references therein).

The BFGS correction is introduced by adding to the local penalized model $\tilde{f}^{(n)}(\boldsymbol{\theta})$ defined in (4.34), the BFGS Hessian proxy $H^{(n)}$ iteratively defined by:

$$H^{(n)} = H^{(n-1)} + \frac{y^{(n-1)} y^{(n-1)\,T}}{y^{(n-1)\,T} s^{(n-1)}} - \frac{H^{(n-1)} s^{(n-1)} s^{(n-1)\,T} H^{(n-1)}}{s^{(n-1)\,T} H^{(n-1)} s^{(n-1)}}.$$

with $H^{(0)}$ an arbitrary positive semidefinite matrix and where $s^{(n-1)} := \boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}^{(n-1)}$ and $y^{(n-1)} := \nabla f(\boldsymbol{\theta}^{(n)}) - \nabla f(\boldsymbol{\theta}^{(n-1)})$. More specifically, we replace the local penalized model $\tilde{f}^{(n)}(\boldsymbol{\theta})$ by

$$\hat{f}^{(n)}(\boldsymbol{\theta}) := \tilde{f}^{(n)}(\boldsymbol{\theta}) + \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(n)}\right)^T H^{(n)} \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(n)}\right),$$

whose gradient is obviously given by:

$$\hat{g}^{(n)}(\boldsymbol{\theta}) := \nabla \hat{f}^{(n)}(\boldsymbol{\theta}) = \nabla \tilde{f}^{(n)}(\boldsymbol{\theta}) + H^{(n)} \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(n)}\right) = \widetilde{g}^{(n)}(\boldsymbol{\theta}) + H^{(n)} \left(\boldsymbol{\theta} - \boldsymbol{\theta}^{(n)}\right),$$

with $\widetilde{g}^{(n)}(\boldsymbol{\theta}) = \nabla \tilde{f}^{(n)}(\boldsymbol{\theta})$ given by (4.36)–(4.38). Using this corrected local penalized model, the solution of the optimization problem will be obtained by iteratively computing

$$\boldsymbol{\theta}^{(n+1)} = \underset{\boldsymbol{\theta} \in \mathbb{R}^N \times \mathbb{M}_N \times \mathbb{M}_N,}{arg\,min} \hat{f}^{(n)}(\boldsymbol{\theta}). \tag{4.63}$$

This is carried out by finding the solution $\boldsymbol{\theta}_0$ of the equation

$$\hat{g}^{(n)}(\boldsymbol{\theta}_0) = \widetilde{g}^{(n)}(\boldsymbol{\theta}_0) + H^{(n)} \left(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^{(n)}\right) = 0. \tag{4.64}$$

using a modified version of the Newton-Raphson iterative scheme spelled out in (**??**). Indeed, it is easy to show that $\boldsymbol{\theta}_0$ is the limit of the sequence $\{\boldsymbol{\theta}^{(n,\,k)}\}_{k \in \mathbb{N}}$ constructed exactly as in Section **??** where the linear systems (**??**) are replaced by

$$\left(\widetilde{T_{\boldsymbol{\theta}^{(n,\,k)}} g^{(n)}} + \widetilde{H^{(n)}}\right) \cdot \widetilde{\boldsymbol{\theta}^{(n,\,k+1)}} = -vec\left(\nabla \tilde{f}^{(n)}(\boldsymbol{\theta}^{(n,\,k)})\right) + \widetilde{H^{(n)}} \cdot \widetilde{\boldsymbol{\theta}^{(n)}} + \widetilde{T_{\boldsymbol{\theta}^{(n,\,k)}} g^{(n)}} \cdot \widetilde{\boldsymbol{\theta}^{(n,\,k)}}. \tag{4.65}$$

where $\widetilde{\boldsymbol{\theta}^{(n,\,k+1)}} = \begin{pmatrix} vec(A^{(n,\,k+1)}) \\ vec(B^{(n,\,k+1)}) \\ \mathbf{c}^{(n,\,k+1)} \end{pmatrix}$ and $\widetilde{H^{(n)}} \in \mathbb{M}_{2N^2+N}$ denotes the matrix associated to $H^{(n)}$ that satisfies

$$vec\left(H^{(n)} \cdot \boldsymbol{\theta}\right) = \widetilde{H^{(n)}} \cdot \begin{pmatrix} vec(A) \\ vec(B) \\ \mathbf{c} \end{pmatrix} \quad \text{for any} \quad \boldsymbol{\theta} = (A, B, \mathbf{c}).$$

Important remark: the Newton-Raphson method and the constraints. In Section 4 we explained how the use of Bregman divergences ensures that at each iteration, the extremum of the local penalized model satisfies the constraints of the problem. However, the

implementation of the Newton-Raphson method that provides the root of the equation (4.64) does not, in general, respect the constraints, and hence this point requires especial care.

In the construction of our optimization algorithm we have used the following prescription in order to ensure that all the elements of the sequence $\{\boldsymbol{\theta}^{(n,\,k)}\}_{k\in\mathbb{N}}$ that converge to the root $\boldsymbol{\theta}_0$ satisfy the constraints: given $\boldsymbol{\theta}^{(n,\,1)}=\boldsymbol{\theta}^{(n)}$ (that satisfies the constraints) let $\boldsymbol{\theta}^{(n,\,2)}$ be the second value in the Newton-Raphson sequence obtained by solving the linear system (4.65). If the value $\boldsymbol{\theta}^{(n,\,2)}$ thereby constructed satisfies the constraints it is then accepted and we continue to the next iteration; otherwise we set

$$\boldsymbol{\theta}^{(n,\,2)}:=\boldsymbol{\theta}^{(n,\,1)}+\frac{\boldsymbol{\theta}^{(n,\,2)}-\boldsymbol{\theta}^{(n,\,1)}}{2} \tag{4.66}$$

iteratively until $\boldsymbol{\theta}^{(n,\,2)}$ satisfies the constraints. Notice that by repeatedly performing (4.66), the value $\boldsymbol{\theta}^{(n,\,2)}$ hence constructed is closer and closer to $\boldsymbol{\theta}^{(n,\,1)}$; since this latter point satisfies the constraints, so will at some point $\boldsymbol{\theta}^{(n,\,2)}$. This manipulation that took us from $\boldsymbol{\theta}^{(n,\,1)}$ to $\boldsymbol{\theta}^{(n,\,2)}$ in a constraint compliant fashion has to be carried out at each iteration to go from $\boldsymbol{\theta}^{(n,\,k)}$ to $\boldsymbol{\theta}^{(n,\,k+1)}$.

Trust-region correction: given an starting point $\boldsymbol{\theta}^0$ we have given a prescription for the construction of a sequence $\{\boldsymbol{\theta}^{(n)}\}_{n\in\mathbb{N}}$ that converges to the constrained minimum of minus the quasi-loglikelihood $f(\boldsymbol{\theta}):=-logL(\mathbf{z};\boldsymbol{\theta})$. We now couple this optimization routine with a trust-region technique. The trust-region algorithm provides us with a systematic method to test the pertinence of an iteration before it is accepted and to adaptively modify the strength of the local penalization in order to speed up the convergence speed. In order to carefully explain our use of this procedure consider first the local model (4.33) in which all the constants $L_1,\ldots,L_6$ that manage the strength of the constraint penalizations are set to a common value L. At each iteration of (4.63) compute the **adequacy ratio** $\rho^{(n)}$ defined as

$$\rho^{(n)}:=\frac{f(\boldsymbol{\theta}^{(n)})-f(\boldsymbol{\theta}^{(n-1)})}{\hat{f}^{(n)}(\boldsymbol{\theta}^{(n)})-\hat{f}^{(n)}(\boldsymbol{\theta}^{(n-1)})}$$

which measures how close the descent in the target function in the present iteration is to the one exhibited by the local model $\hat{f}^{(n)}$. The values that can be obtained for $\rho^{(n)}$ are classified into three categories that determine different courses of action:

(a) **Bad iteration** $\rho^{(n)}<0.01$: there is too much dissimilarity between the local penalized model and the actual target function. In this situation, the iteration update is rejected by setting $\boldsymbol{\theta}^{(n)}=\boldsymbol{\theta}^{(n-1)}$ and the penalization is strengthened by doubling the constant: $L=2L$

(b) **Good iteration** $0.01\leqslant\rho^{(n)}\leqslant 0.9$: the iteration update is accepted and the constant L is left unchanged.

(c) **Very good iteration** $0.9\leqslant\rho^{(n)}$: the iteration update is accepted but given the very good adequacy between the local penalized model and the target function we can afford loosening the penalization strength by setting $L=\frac{1}{2}L$ as the constant that will be used in the next iteration.

Preliminary estimation

As any optimization algorithm, the one that we just presented requires a starting point $\boldsymbol{\theta}^{(0)}$. The choice of a good preliminary estimation of $\boldsymbol{\theta}^{(0)}$ is particularly relevant in our situation

since the quasi-loglikelihood exhibits generically local extrema and hence initializing the optimization algorithm close enough to the solution may prove to be crucial in order to obtain the correct solution.

Given a sample $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_T\}$, a reasonable estimation for $\boldsymbol{\theta}^{(0)}$ can be obtained by using the following two steps scheme:

1. Find a preliminary estimation of the conditional covariance matrices sequence $\{H_1, \dots, H_T\}$ out of the sample $\mathbf{z}$. This can be achieved by using a variety of existing non-computationally intensive techniques. A non-exhaustive list is:

(i) Orthogonal GARCH model (O-GARCH): introduced in [**?**, **?**, **?**, **?**]; this technique is based on fitting one-dimensional GARCH models to the principal components obtained out of the sample marginal covariance matrix of $\mathbf{z}$.

(ii) Generalized orthogonal GARCH model (GO-GARCH) [**?**]: similar to O-GARCH, but in this case the one-dimensional modeling is carried out not for the principal components of $\mathbf{z}$ but for its image with respect to a transformation V which is assumed is assumed to be just invertible (in the case of O-GARCH is also orthogonal) and it is estimated directly via a maximum likelihood procedure, together with the parameters of the one-dimensional GARCH models. GO-GARCH produces better empirical results than O-GARCH but it lacks the factoring calibration feature that O-GARCH has, making it more complicated for the modeling of large dimensional time series and conditional covariance matrices.

(iii) Independent component analysis (ICA-GARCH): [**?**] this model is based on a signal separation technique [**?**, **?**] that turns the time series into statistically independent components that are then treated separately using one dimensional GARCH models.

(iv) Dynamic conditional correlation model (DCC): introduced in [**?**, **?**], this model proposes a dynamic behavior of the conditional correlation that depends on a small number of parameters and that nevertheless is still capable of capturing some of the features of more complicated multivariate models. Moreover, a version of this model [**?**] can be estimated consistently using a two-step approach that makes it suitable to handle large dimensional problems.

Other method that is widely used in the context of financial log-returns is the one advocated by Riskmetrics [**?**] that proposes exponentially weighted moving average (EWMA) models for the time evolution of variances and covariances; this comes down to working with IGARCH type models with a coefficient that is not estimated but proposed by Riskmetrics and that is the same for all the factors.

2. Estimation of $\boldsymbol{\theta}^{(0)}$ out of $\mathbf{z}$ and $H = \{H_t\}_{t\in\{1,\dots,T\}}$ using constrained ordinary least squares. If we have the sample $\mathbf{z}$ and a preliminary estimation of the conditional covariances $\{H_t\}_{t\in\{1,\dots,T\}}$, a good candidate for $\boldsymbol{\theta}^{(0)} = (A^{(0)}, B^{(0)}, \mathbf{c}^{(0)})$ is the value that minimizes the sum of the Euclidean norms $s_t := \|\mathbf{h}_t - \left(\mathbf{c} + A\boldsymbol{\eta}_{t-1} + B\mathbf{h}_{t-1}\right)\|^2$, that is,

$$s(A, B, \mathbf{c}; \mathbf{z}, H) = \sum_{t=2}^{T} s_t(A, B, \mathbf{c}; \mathbf{z}, H) = \sum_{t=2}^{T} \|\mathbf{h}_t - \left(\mathbf{c} + A\boldsymbol{\eta}_{t-1} + B\mathbf{h}_{t-1}\right)\|^2,$$

subjected to the constraints **(SC)**, **(PC)**, **(CC)**, and **(KC)**. This minimum can be efficiently found by using the Bregman divergences based method introduced in Sections 4 through 4 with the function $s(A, B, \mathbf{c}; \mathbf{z}, H)$ replacing minus the log-likelihood. However, we emphasize

that unlike the situation in the log-likelihood problem, the choice of a starting point in the optimization of $s(A, B, \mathbf{c}; \mathbf{z}, H)$ is irrelevant given the convexity of his function.

As a consequence of these arguments, the preliminary estimation $\boldsymbol{\theta}^{(0)}$ is obtained by iterating (4.63) where in the local model (4.34) the map f is replaced by s. This scheme is hence readily applicable once the gradient of s, provided by the following formulas, is available:

$$\begin{aligned}
\nabla_A s &= 2\sum_{t=2}^{T}\left[A\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^T + \mathbf{c}\boldsymbol{\eta}_{t-1}^T + B\mathbf{h}_{t-1}\boldsymbol{\eta}_{t-1}^T - \mathbf{h}_t\boldsymbol{\eta}_{t-1}^T\right], \\
\nabla_B s &= 2\sum_{t=2}^{T}\left[\mathbf{c}\mathbf{h}_{t-1}^T + A\boldsymbol{\eta}_{t-1}\mathbf{h}_{t-1}^T + B\mathbf{h}_{t-1}\mathbf{h}_{t-1}^T - \mathbf{h}_t\mathbf{h}_{t-1}^T\right], \\
\nabla_{\mathbf{c}} s &= 2\sum_{t=2}^{T}\left[\mathbf{c} + A\boldsymbol{\eta}_{t-1} + B\mathbf{h}_{t-1} - \mathbf{h}_t\right].
\end{aligned}$$

5 Appendix

Proof of Proposition 2.1

We start with the proof of **(i)** by using the following chain of equalities in which we use the symmetric character of both A and $math(m)$:

$$\begin{aligned}
\langle A + diag(A), math(m)\rangle &= \operatorname{trace}(A\, math(m)) + \operatorname{trace}(diag(A)math(m)) \\
&= \sum_{i,j=1}^{n} A_{ij}math(m)_{ji} + A_{ij}\delta_{ij}math(m)_{ji} \\
&= \sum_{i<j} A_{ij}math(m)_{ij} + \sum_{i>j} A_{ij}math(m)_{ij} + 2\sum_{i=j=1}^{n} A_{ij}math(m)_{ij} \\
&= 2\sum_{i\geqslant j} A_{ij}math(m)_{ij} = 2\sum_{i\geqslant j} A_{ij}m_{\sigma(i,j)} = 2\sum_{q=1}^{N} A_{\sigma^{-1}(q)}m_q \\
&= 2\langle vech(A), m\rangle,
\end{aligned}$$

as required. In order to prove **(ii)**, note that the identity that we just showed ensures that

$$\langle A, math(m)\rangle = 2\langle vech(A), m\rangle - \langle diag(A), math(m)\rangle. \tag{5.67}$$

At the same time

$$\begin{aligned}
\langle diag(A), math(m)\rangle &= \operatorname{trace}(diag(A)math(m)) = \sum_{i=1}^{n} A_{ii}math(m)_{ii} = \sum_{i=1}^{n} A_{ii}m_{\sigma(i,i)} \\
&= \sum_{i\geqslant j} diag(A)_{ij}m_{\sigma(i,j)} = \sum_{q=1}^{N} diag(A)_{\sigma^{-1}(q)}m_q \\
&= \sum_{q=1}^{N} vech(diag(A))_q m_q = \langle vech(diag(A)), m\rangle,
\end{aligned}$$

which substituted in the right hand side of (5.67) proves the required identity. Finally, expression (2.3) follows directly from **(ii)** and as to (2.4) we observe that

$$\begin{aligned}\frac{1}{2}\langle A + diag(A), math(m)\rangle &= \frac{1}{2}\operatorname{trace}((A + diag(A))math(m)) \\ &= \frac{1}{2}\operatorname{trace}(Amath(m)) + \frac{1}{2}\operatorname{trace}(diag(A)math(m)) \\ &= \frac{1}{2}\left(\operatorname{trace}(Amath(m)) + \operatorname{trace}(Adiag(math(m)))\right) \\ &= \frac{1}{2}\langle A, math(m) + diag(math(m))\rangle,\end{aligned}$$

which proves (2.4). Regarding the operator norms we will just prove (2.5) and (2.6) as the rest can be easily obtained out of these two combined with the expressions (2.3) and (2.4). We start by noticing that for any nonzero $A = (a_{ij}) \in \mathbb{S}_n$:

$$\frac{\|vech(A)\|^2}{\|A\|^2} = \frac{\sum_{i>j=1}^n a_{ij}^2 + \sum_{i=1}^n a_{ii}^2}{2\sum_{i>j=1}^n a_{ij}^2 + \sum_{i=1}^n a_{ii}^2} = 1 - \frac{\sum_{i>j=1}^n a_{ij}^2}{2\sum_{i>j=1}^n a_{ij}^2 + \sum_{i=1}^n a_{ii}^2}.$$

Since the last summand in the previous expression is always positive we have that

$$\|vech\|_{op} = \sup_{A\in\mathbb{S}_n, A\neq 0} \frac{\|vech(A)\|}{\|A\|} = 1,$$

the supremum being attained by any diagonal matrix ($\sum_{i>j=1}^n a_{ij}^2 = 0$ in that case). Consider now $v = vech(A)$. Then:

$$\frac{\|math(v)\|^2}{\|v\|^2} = \frac{\|A\|^2}{\|vech(A)\|^2} = \frac{2\sum_{i>j=1}^n a_{ij}^2 + \sum_{i=1}^n a_{ii}^2}{\sum_{i>j=1}^n a_{ij}^2 + \sum_{i=1}^n a_{ii}^2} = 1 + \frac{\sum_{i>j=1}^n a_{ij}^2}{\sum_{i>j=1}^n a_{ij}^2 + \sum_{i=1}^n a_{ii}^2}. \tag{5.68}$$

When we let $A \in \mathbb{S}_n$ vary in the previous expression, we obtain a supremum by considering matrices with zeros in the diagonal ($\sum_{i=1}^n a_{ii}^2 = 0$) and by choosing $\sum_{i>j=1}^n a_{ij}^2 \to \infty$, in which case $\frac{\|A\|^2}{\|vech(A)\|^2} \to 2$. Finally, as the map $vech : \mathbb{S}_n \to \mathbb{R}^N$ is an isomorphism, (5.68) implies that

$$\|math\|_{op} = \sup_{v\in\mathbb{R}^N, v\neq 0} \frac{\|math(v)\|}{\|v\|} = \sup_{A\in\mathbb{S}_n, A\neq 0} \frac{\|A\|}{\|vech(A)\|} = \sqrt{2}. \quad \blacksquare$$

Proof of Proposition 2.3

We just need to verify that (2.11) satisfies (2.12). Let $k, l \in \{1, \dots, n\}$ be such that $k \geqslant l$. Then,

$$\begin{aligned}(A\,vech(H))_{\sigma(k,l)} &= \sum_{i\geqslant j} A_{\sigma(k,l),\sigma(i,j)} H_{ij} = \sum_{i\geqslant j} A_{\sigma(k,l),\sigma(i,j)} \frac{H_{ij} + H_{ji}}{2} \\ &= \frac{1}{2}\sum_{i\geqslant j} A_{\sigma(k,l),\sigma(i,j)} H_{ij} + \frac{1}{2}\sum_{i\geqslant j} A_{\sigma(k,l),\sigma(i,j)} H_{ji} \\ &= \frac{1}{2}\sum_{i>j} A_{\sigma(k,l),\sigma(i,j)} H_{ij} + \sum_{i=j} A_{\sigma(k,l),\sigma(i,j)} H_{ij} + \frac{1}{2}\sum_{i<j} A_{\sigma(k,l),\sigma(j,i)} H_{ij} \\ &= \sum_{i>j} (\Sigma(A)_{kl})_{ij} H_{ij} + \sum_{i=j} (\Sigma(A)_{kl})_{ij} H_{ij} + \sum_{i<j} (\Sigma(A)_{kl})_{ij} H_{ij} = \operatorname{trace}(\Sigma(A)_{kl} H),\end{aligned}$$

as required. $\blacksquare$

Proof of Proposition 2.4

We start with the following Lemma:

Lemma 5.1 *Let $A \in \mathbb{M}_{n^2}$. The orthogonal projections $\mathbb{P}_{n^2}(A) \in \mathbb{S}_{n^2}$ and $\mathbb{P}^n_{n^2}(A) \in \mathbb{S}^n_{n^2}$ of A onto the spaces of symmetric and n-symmetric matrices with respect to the Frobenius inner product (2.1) are given by:*

$$\mathbb{P}_{n^2}(A) = \frac{1}{2}(A + A^T) \tag{5.69}$$

$$(\mathbb{P}^n_{n^2}(A))_{kl} = \frac{1}{4}(A_{kl} + A^T_{kl} + A_{lk} + A^T_{lk}), \tag{5.70}$$

for any block $(\mathbb{P}^n_{n^2}(A))_{kl}$ of $\mathbb{P}^n_{n^2}(A)$, $k, l \in \{1, \dots, n\}$.

Proof. In order to prove (5.69) it suffices to check that $\langle A - \mathbb{P}_{n^2}(A), B\rangle = 0$ for any $B \in \mathbb{S}_{n^2}$. Indeed,

$$\langle A - \mathbb{P}_{n^2}(A), B\rangle = \operatorname{trace}(AB) - \frac{1}{2}\operatorname{trace}(AB) - \frac{1}{2}\operatorname{trace}(A^T B) = 0.$$

The result follows from the uniqueness of the orthogonal projection. Regarding (5.70) we check that $\langle A - \mathbb{P}^n_{n^2}(A), B\rangle = 0$, for any $B \in \mathbb{S}^n_{n^2}$. Given that for any $k, l \in \{1, \dots, n\}$ the block $(AB)_{kl}$ is given by $(AB)_{kl} = \sum_{r=1}^n A_{kr}B_{rl}$ we have

$$\begin{aligned}
\langle A - \mathbb{P}^n_{n^2}(A), B\rangle &= \operatorname{trace}(AB) - \operatorname{trace}(\mathbb{P}^n_{n^2}(A)B) = \sum_{i=1}^n \operatorname{trace}(AB)_{ii} - \operatorname{trace}(\mathbb{P}^n_{n^2}(A)B)_{ii} \\
&= \sum_{i,j=1}^n \operatorname{trace}(A_{ij}B_{ji}) - \operatorname{trace}((\mathbb{P}^n_{n^2}(A))_{ij}B_{ji}) = \sum_{i,j=1}^n \operatorname{trace}(A_{ij}B_{ji}) \\
&- \sum_{i,j=1}^n \left[\frac{1}{4}\operatorname{trace}(A_{ij}B_{ji}) + \frac{1}{4}\operatorname{trace}(A^T_{ij}B_{ji}) + \frac{1}{4}\operatorname{trace}(A_{ji}B_{ji}) + \frac{1}{4}\operatorname{trace}(A^T_{ji}B_{ji})\right] = 0,
\end{aligned}$$

where we used that, due to the n-symmetricity of B $\operatorname{trace}(A^T_{ij}B_{ji}) = \operatorname{trace}(B^T_{ji}A_{ij}) = \operatorname{trace}(A_{ij}B_{ji})$ and

$$\sum_{i,j=1}^n \operatorname{trace}(A_{ji}B_{ji}) = \operatorname{trace}(A_{ji}B_{ij}) = \operatorname{trace}(A_{ij}B_{ij}).$$

Analogously $\sum_{i,j=1}^n \operatorname{trace}(A^T_{ji}B_{ji}) = \operatorname{trace}(A_{ij}B_{ij})$. ■

Now, in order to prove Proposition 2.4, consider $A \in \mathbb{M}_N$ and $\mathcal{B} \in \mathbb{M}_{n^2}$. Since the image of the map Σ lies in $\mathbb{S}^2_{n^2}$ we have that $\langle \mathcal{B} - \mathbb{P}^n_{n^2}(\mathcal{B}), \Sigma(A)\rangle = 0$ and hence

$$\langle \Sigma^*(\mathcal{B}), A\rangle = \langle \mathcal{B}, \Sigma(A)\rangle = \langle \mathbb{P}^n_{n^2}(\mathcal{B}) + \mathcal{B} - \mathbb{P}^n_{n^2}(\mathcal{B}), \Sigma(A)\rangle = \langle \mathbb{P}^n_{n^2}(\mathcal{B}), \Sigma(A)\rangle = \langle \Sigma^*(\mathbb{P}^n_{n^2}(\mathcal{B})), A\rangle.$$

This identity allows us to restrict the proof of (2.4) to the n-symmetric elements $\mathcal{B} \in \mathbb{S}^n_{n^2}$. Hence let $\mathcal{B} \in \mathbb{S}^n_{n^2}$ and let $\widetilde{\sigma}$ be the extension of the map σ defined in (2.2). Then,

$$\begin{aligned}
\langle \Sigma(A), \mathcal{B} \rangle &= \sum_{k,l=1}^{n} \langle \Sigma(A)_{kl}, \mathcal{B}_{kl} \rangle = \sum_{k,l=1}^{n} \operatorname{trace}(\Sigma(A)_{kl} \mathcal{B}_{kl}^T) = \sum_{k,l,i,j=1}^{n} (\Sigma(A)_{kl})_{ij} (\mathcal{B}_{kl})_{ij} \\
&= \sum_{k,l,i,j=1}^{n} \frac{1}{2} \left[A_{\widetilde{\sigma}(k,l),\widetilde{\sigma}(i,j)} + A_{\widetilde{\sigma}(k,l),\widetilde{\sigma}(i,j)} \delta_{ij} \right] (\mathcal{B}_{kl})_{ij} \\
&= \sum_{k,j=1}^{n} \left[\sum_{i<j}^{n} \frac{1}{2} A_{\widetilde{\sigma}(k,l),\sigma(j,i)} (\mathcal{B}_{kl})_{ji} + \sum_{i=j=1}^{n} A_{\widetilde{\sigma}(k,l),\sigma(i,j)} (\mathcal{B}_{kl})_{ij} + \frac{1}{2} \sum_{i>j}^{n} A_{\widetilde{\sigma}(k,l),\sigma(i,j)} (\mathcal{B}_{kl})_{ij} \right] \\
&= \sum_{k,j=1}^{n} \sum_{i \geqslant j}^{n} A_{\widetilde{\sigma}(k,l),\sigma(i,j)} (\mathcal{B}_{kl})_{ji} \\
&= \sum_{i \geqslant j}^{n} \left[\sum_{k<l}^{n} A_{\sigma(l,k),\sigma(j,i)} (\mathcal{B}_{lk})_{ji} + \sum_{k=l=1}^{n} A_{\sigma(k,l),\sigma(i,j)} (\mathcal{B}_{kl})_{ij} + \sum_{l<k}^{n} A_{\sigma(k,l),\sigma(i,j)} (\mathcal{B}_{kl})_{ij} \right] \\
&= \sum_{i \geqslant j}^{n} \left[\sum_{k \geqslant l}^{n} A_{\sigma(k,l),\sigma(i,j)} (\mathcal{B}_{kl})_{ij} - \sum_{k=l=1}^{n} A_{\sigma(k,l),\sigma(i,j)} (\mathcal{B}_{kl})_{ij} \delta_{kl} \right] \\
&= \sum_{p,q=1}^{N} \left[2A_{p,q} B_{p,q} - A_{p,q} B_{p,q} \delta_{pr_1(\sigma^{-1}(p)),pr_2(\sigma^{-1}(p))} \right] = \operatorname{trace}(2AB^T - A\widetilde{B}^T) = \langle A, 2B -
\end{aligned}$$

which proves the statement. We emphasize that in the fourth and sixth equalities we used the n-symmetry of $\mathcal{B}$. The equality (2.14) is proved in a straightforward manner by verifying that $\widetilde{\Sigma}^{-1} \circ \Sigma = \mathbb{I}_{\mathbb{M}_N}$ and $\Sigma \circ \widetilde{\Sigma}^{-1} = \mathbb{I}_{\mathbb{S}^n_{n^2}}$ using the defining expressions (2.2) and (2.14). ■

Proof of Proposition 3.1

Using the property of the operator Σ stated in Proposition 2.3, the second equality in (3.15) can be rewritten as:

$$\begin{aligned}
vech(H_t) &= vech(math(\mathbf{c})) + Avech(\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T) + Bvech(H_{t-1}) \\
&= vech(math(\mathbf{c})) + vech(\Sigma(A) \bullet (\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T)) + vech(\Sigma(B) \bullet H_{t-1}),
\end{aligned}$$

or, equivalently:

$$H_t = math(\mathbf{c}) + \Sigma(A) \bullet (\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T) + \Sigma(B) \bullet H_{t-1}.$$

In view of this expression and in the terms of the statement of the proposition, it suffices to show that both $\Sigma(A) \bullet (\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T)$ and $\Sigma(B) \bullet H_{t-1}$ are positive semidefinite provided that H_{t-1} is positive semidefinite. Regarding $\Sigma(A) \bullet (\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T)$, consider $\mathbf{v} \in \mathbb{R}^{n^2}$. Then

$$\begin{aligned}
\langle \mathbf{v}, \Sigma(A) \bullet (\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T)\mathbf{v} \rangle &= \sum_{i,j=1}^{n^2} v_i (\Sigma(A) \bullet (\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T))_{ij} v_j = \sum_{i,j=1}^{n^2} v_i \operatorname{trace}(\Sigma(A)_{ij} (\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T)) v_j \\
&= \sum_{i,j=1}^{n^2} v_i \operatorname{trace}(\mathbf{z}_{t-1}^T \Sigma(A)_{ij} \mathbf{z}_{t-1}) v_j = \sum_{i,j,k,l=1}^{n^2} v_i z_{t-1,k}^T (\Sigma(A)_{ij})_{kl} z_{t-1,l} v_j \\
&= \langle \mathbf{v} \otimes \mathbf{z}_{t-1}, \Sigma(A)(\mathbf{v} \otimes \mathbf{z}_{t-1}) \rangle,
\end{aligned}$$

which is greater or equal to zero due to the positive semidefiniteness hypothesis on $\Sigma(A)$. In the last equality we used (2.10).

As to $\Sigma(B) \bullet H_{t-1}$, we start by noticing that $H_{t-1} = E_{t-1}[\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T]$ and hence $\Sigma(B) \bullet H_{t-1} = \Sigma(B) \bullet E_{t-1}[\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T]$. This equality, as well as the linearity of the conditional expectation allows us to use virtually the same argument as above. Indeed, for any $\mathbf{v} \in \mathbb{R}^{n^2}$

$$\begin{aligned}\langle \mathbf{v}, \Sigma(B) \bullet H_{t-1}\mathbf{v}\rangle &= \sum_{i,j=1}^{n^2} v_i \operatorname{trace}(\Sigma(B)_{ij} E_{t-1}[\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T])v_j = \sum_{i,j=1}^{n^2} E_{t-1}[v_i \operatorname{trace}(\Sigma(B)_{ij}\mathbf{z}_{t-1}\mathbf{z}_{t-1}^T)v_j] \\ &= E_{t-1}[\langle \mathbf{v} \otimes \mathbf{z}_{t-1}, \Sigma(B)(\mathbf{v} \otimes \mathbf{z}_{t-1})\rangle],\end{aligned}$$

which is greater or equal to zero due to the positive semidefiniteness hypothesis on $\Sigma(B)$. ■

Proof of Proposition 3.2

We start by noticing that the VEC(1,1) model is by construction a white noise and hence it suffices to establish the stationarity of the variance. Indeed, for any $t, h \in \mathbb{N}$ we compute the autocovariance function Γ:

$$\begin{aligned}\Gamma(t, t+h) := E\left[\mathbf{z}_t\mathbf{z}_{t+h}^T\right] &= E\left[E_t\left[H_t^{1/2}\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_{t+h}}H_{t+h}^{1/2}\right]\right] \\ &= E\left[H_t^{1/2}E_t\left[\boldsymbol{\varepsilon_t}\boldsymbol{\varepsilon_{t+h}}\right]H_{t+h}^{1/2}\right] = \delta_{h0}E\left[H_t^{1/2}H_{t+h}^{1/2}\right]. \quad (5.71)\end{aligned}$$

Consequently, we just need to prove the existence of a solution for which $\Gamma(t,t) = E[H_t]$ or, equivalently $E[\mathbf{h}_t]$, is time independent. We first notice that

$$E[\mathbf{h}_t] = E\left[\mathbf{c} + A\boldsymbol{\eta}_{t-1} + B\boldsymbol{h}_{t-1}\right] = E\left[\mathbf{c} + A\boldsymbol{h}_{t-1} + B\boldsymbol{h}_{t-1}\right] + A\,E\left[\boldsymbol{\eta}_{t-1} - \boldsymbol{h}_{t-1}\right] = E\left[\mathbf{c} + A\boldsymbol{h}_{t-1} + B\boldsymbol{h}_{t-1}\right],$$

since $A\,E\left[\boldsymbol{\eta}_{t-1} - \boldsymbol{h}_{t-1}\right] = 0$ by (5.71). Now, for any $k > 0$

$$E[\mathbf{h}_t] = \mathbf{c} + (A+B)E\left[\mathbf{h}_{t-1}\right] = \sum_{j=0}^{k}(A+B)^j\mathbf{c} + (A+B)^{k+1}E\left[\mathbf{h}_{t-k-1}\right].$$

If all the eigenvalues of $A + B$ are smaller than one in modulus then (see, for example [**?**, Appendix A.9.1])

$$\sum_{j=0}^{k}(A+B)^j\mathbf{c} \xrightarrow[k\to\infty]{} (\mathbb{I}_N - A - B)^{-1}\mathbf{c}, \qquad \text{and} \qquad (A+B)^{k+1}E\left[\mathbf{h}_{t-k-1}\right] \xrightarrow[k\to\infty]{} 0,$$

in which case $E[\mathbf{h}_t]$ is time independent and

$$\Gamma(0) = math(E[\mathbf{h}_t]) = math((\mathbb{I}_N - A - B)^{-1}\mathbf{c}).$$

The sufficient condition in terms of the top singular value $\sigma_{max}(A+B)$ of $A+B$ is a consequence of the fact that (see for instance [**?**, Theorem 5.6.9]) $|\lambda(A+B)| \leqslant \sigma_{max}(A+B)$, for any eigenvalue $\lambda(A+B)$ of $A+B$. ■

Proof of Proposition 3.3

The chain rule implies that for any perturbation Δ in the θ direction

$$d_\theta l_t \cdot \Delta = d_{H_t} l_t(H_t(\theta)) \cdot T_\theta H_t \cdot \Delta = \langle \nabla_{H_t} l_t, T_\theta H_t \cdot \Delta \rangle = \langle T_\theta^* H_t \cdot \nabla_{H_t} l_t, \Delta \rangle,$$

which proves that $\nabla_\theta l_t = T_\theta^* H_t \cdot \nabla_{H_t} l_t$ and hence (3.22) follows. We now establish (3.23) by showing separately that

$$\nabla_{H_t} \log(\det(H_t)) = H_t^{-1} \quad \text{and} \quad \nabla_{H_t} \left(-\frac{1}{2} \mathbf{z}_t^T H_t^{-1} \mathbf{z}_t \right) = \frac{1}{2} \left(H_t^{-1} \mathbf{z}_t \mathbf{z}_t^T H_t^{-1} \right). \tag{5.72}$$

In order to prove the first expression we start by using the positive semidefinite character of H_t in order to write $H_t = VDV^T$. V is an orthogonal matrix and D is diagonal with non-negative entries; it has hence a unique square root $D^{1/2}$ that we can use to write $H_t = VDV^T = (VD^{1/2})(VD^{1/2})^T$. Let $\delta \in \mathbb{R}$ and $\Delta \in \mathbb{S}_n$. We have

$$\begin{aligned}
\log(\det(H_t + \delta\Delta)) &= \log(\det((VD^{1/2})(VD^{1/2})^T + \delta\Delta)) \\
&= \log(\det((VD^{1/2})(\mathbb{I}_n + \delta(D^{-1/2}V^T)\Delta(VD^{-1/2}))(VD^{1/2})^T)) \\
&= \log(\det(VD^{1/2}) \det(\mathbb{I}_n + \delta(D^{-1/2}V^T)\Delta(VD^{-1/2})) \det(VD^{1/2})^T) \\
&= \log(\det((VD^{1/2})(VD^{1/2})^T) \det(\mathbb{I}_n + \delta(D^{-1/2}V^T)\Delta(VD^{-1/2}))) \\
&= \log(\det(H_t) \det(\mathbb{I}_n + \delta\Xi)),
\end{aligned}$$

with $\Xi := (D^{-1/2}V^T)\Delta(VD^{-1/2})$. This matrix is symmetric and hence normal and diagonalizable; let $\{\lambda_1, \ldots, \lambda_n\}$ be its eigenvalues. We hence have that

$$\begin{aligned}
dH_t \cdot \Delta &= \frac{d}{d\delta}\Big|_{\delta=0} \log(\det(H_t + \delta\Delta)) = \frac{d}{d\delta}\Big|_{\delta=0} \log(\det(H_t)) + \log\left(\prod_{i=1}^n (1 + \delta\lambda_i) \right) = \frac{d}{d\delta}\Big|_{\delta=0} \sum_{i=1}^n \log \\
&= \sum_{i=1}^n \lambda_i = \operatorname{trace}((D^{-1/2}V^T)\Delta(VD^{-1/2})) = \operatorname{trace}((VD^{-1/2})(D^{-1/2}V^T)\Delta) = \operatorname{trace}(H_t^{-1}\Delta)
\end{aligned}$$

which proves $\nabla_{H_t} \log(\det(H_t)) = H_t^{-1}$. Regarding the second expression in (5.72) we define $f(H_t) := -\frac{1}{2}\mathbf{z}_t^T H_t^{-1} \mathbf{z}_t$ and note that

$$\begin{aligned}
df(H_t) \cdot \Delta &= \frac{d}{dt}\Big|_{t=0} -\frac{1}{2}\mathbf{z}_t^T (H_t + t\Delta)^{-1} \mathbf{z}_t = \frac{d}{dt}\Big|_{t=0} -\frac{1}{2}\mathbf{z}_t^T (\mathbb{I}_n + tH_t^{-1}\Delta)^{-1} H_t^{-1} \mathbf{z}_t \\
&= \frac{d}{dt}\Big|_{t=0} -\frac{1}{2}\mathbf{z}_t^T (\mathbb{I}_n + tH_t^{-1}\Delta)^{-1} H_t^{-1} \mathbf{z}_t = \frac{d}{dt}\Big|_{t=0} -\frac{1}{2} \sum_{k=0}^\infty (-1)^k t^k \mathbf{z}_t^T (H_t^{-1}\Delta)^k H_t^{-1} \mathbf{z}_t \\
&= \frac{1}{2}\mathbf{z}_t^T H_t^{-1} \Delta H_t^{-1} \mathbf{z}_t = \frac{1}{2} \operatorname{trace}(H_t^{-1} \mathbf{z}_t \mathbf{z}_t^T H_t^{-1} \Delta),
\end{aligned}$$

which implies that $\nabla_{H_t} f = \frac{1}{2}\left(H_t^{-1} \mathbf{z}_t \mathbf{z}_t^T H_t^{-1} \right)$, as required.

In order to prove (3.24)–(3.26) we notice that the second equation in (3.15) can be rewritten using the vech and math operators as

$$H_t = math\left(\mathbf{c} + A\boldsymbol{\eta}_{t-1} + Bvech(H_{t-1}) \right). \tag{5.73}$$

We now show (3.24). Let $\mathbf{v} \in \mathbb{R}^N$ and $\Delta \in \mathbb{S}_n$ arbitrary. Identity (5.73) and the linearity of the various mappings involved imply that $T_{\mathbf{c}} H_t \cdot \mathbf{v} = math\left(\mathbf{c} + A\boldsymbol{\eta}_{t-1} + Bvech(T_{\mathbf{c}} H_{t-1} \cdot \mathbf{v}) \right)$ and hence

$$\begin{aligned}
\langle T_{\mathbf{c}}^* H_t \cdot \Delta, \mathbf{v} \rangle &= \langle \Delta, T_{\mathbf{c}} H_t \cdot \mathbf{v} \rangle = \langle \Delta, math\left(\mathbf{c} + A\boldsymbol{\eta}_{t-1} + Bvech(T_{\mathbf{c}} H_{t-1} \cdot \mathbf{v}) \right) \rangle \\
&= \langle math^*(\Delta) + T_{\mathbf{c}}^* H_{t-1} \cdot vech^*(B^T math^*(\Delta)), \mathbf{v} \rangle.
\end{aligned}$$

The proof of (3.25) follows a similar scheme. By (5.73) we have that for any $M \in \mathbb{M}_N$:

$$T_A H_t \cdot M = math\left(M\boldsymbol{\eta}_{t-1} + Bvech(T_A H_{t-1} \cdot M)\right). \tag{5.74}$$

Consequently, for any $\Delta \in \mathbb{S}_n$

$$\begin{aligned}\langle T_A^* H_t \cdot \Delta, M\rangle &= \langle \Delta, T_A H_t\rangle = \langle \Delta, math\left(M\boldsymbol{\eta}_{t-1} + Bvech(T_A H_{t-1} \cdot M)\right)\rangle \\ &= \langle math^*(\Delta)\cdot \boldsymbol{\eta}_{t-1}^T + T_A^* H_{t-1} \cdot vech^*(B^T math^*(\Delta)), \Delta\rangle.\end{aligned}$$

Finally, (3.26) is proved analogously replacing (5.74) by its B counterpart, namely,

$$T_B H_t \cdot M = math\left(Mvech(H_{t-1}) + Bvech(T_B H_{t-1} \cdot M)\right). \quad \blacksquare$$

Proof of Proposition 3.4

An inductive argument using (3.24)–(3.26) guarantees that for any $t, k \in \mathbb{N}$, $k \leqslant t$

$$T_{\mathbf{c}} H_t^* \cdot \Delta = \sum_{i=1}^{k} B^{i-1\,T} math^*(\Delta) + T_{\mathbf{c}}^* H_{t-k} \cdot vech^*(B^{k\,T} math^*(\Delta)), \tag{5.75}$$

$$T_A^* H_t \cdot \Delta = \sum_{i=1}^{k} B^{i-1\,T} math^*(\Delta)\cdot \boldsymbol{\eta}_{t-i}^T + T_A^* H_{t-k} \cdot vech^*(B^{k\,T} math^*(\Delta)), \tag{5.76}$$

$$T_B^* H_t \cdot \Delta = \sum_{i=1}^{k} B^{i-1\,T} math^*(\Delta)\cdot vech(H_{t-i})^T + T_B^* H_{t-k} \cdot vech^*(B^{k\,T} math^*(\Delta)), \tag{5.77}$$

The first expression with $k = t$ and the norm estimate (2.8) imply that

$$\|T_{\mathbf{c}} H_t^* \cdot \Delta\| = \left\|\sum_{i=1}^{t} B^{i-1\,T} math^*(\Delta)\right\| \leqslant \sqrt{2}\sum_{i=1}^{t}\|B\|_{op}^{i-1}\|\Delta\| \leqslant \frac{\sqrt{2}\|\Delta\|}{1-\|B\|_{op}}. \tag{5.78}$$

We now use (5.75) for an arbitrary k as well as (2.7) and (5.78) and write

$$\begin{aligned}\|(T_{\mathbf{c}}^* H_t - T_{\mathbf{c}}^* H_t^k)\cdot\Delta\| &= \|T_{\mathbf{c}}^* H_{t-k} \cdot vech^*(B^{k\,T} math^*(\Delta))\| \\ &\leqslant \|T_{\mathbf{c}}^* H_{t-k}\|_{op}\|vech^*\|_{op}\|B\|_{op}^k\|math^*\|_{op}\|\Delta\| \leqslant \frac{2\|\Delta\|\|B\|_{op}^k}{1-\|B\|_{op}}.\end{aligned} \tag{5.79}$$

The computability constraint **(CC)** implies that $\|B\|_{op} \leqslant 1-\widetilde{\varepsilon}_B$ and hence $\|T_{\mathbf{c}}^* H_t - T_{\mathbf{c}}^* H_t^k\|_{op} \leqslant 2(1-\widetilde{\varepsilon}_B)^k/\widetilde{\varepsilon}_B$. A straightforward computation shows that if we want this upper bound for the error to be smaller than a certain $\delta > 0$, that is $2(1-\widetilde{\varepsilon}_B)^k/\widetilde{\varepsilon}_B < \delta$ then it suffices to take

$$k > \frac{\log\left(\frac{\widetilde{\varepsilon}_B\delta}{2}\right)}{\log(1-\widetilde{\varepsilon}_B)}. \tag{5.80}$$

We now tackle the estimation of the truncation error in mean in the A variable. Firstly, we recall that by (5.71) and in the presence of the stationarity constraint $E[\boldsymbol{\eta}_t] = E[\mathbf{h}_t] = (\mathbb{I}_N - A - B)^{-1}\mathbf{c}$. The first consequence of this identity is that if we take the expectations of both (5.76) and (5.77) we see that $\|E\left[T_A^* H_t \cdot \Delta\right]\|$ and $\|E\left[T_B^* H_t \cdot \Delta\right]\|$ are determined by

exactly the same recursions and hence the error estimations for both variables are going to be the same. Also, by (5.76)

$$\|E\left[T_A^* H_t \cdot \Delta\right]\| = \left\| \sum_{i=1}^{t} B^{i-1\,T} math^*(\Delta) \cdot E[\boldsymbol{\eta}_{t-i}^T] \right\| \leqslant \sqrt{2}\|\Delta\|\|E[\mathbf{h}_t]\| \sum_{i=1}^{t} \|B\|_{op}^{i-1}$$
$$\leqslant \sqrt{2}\|\Delta\|\|(\mathbb{I}_N - A - B)^{-1}\mathbf{c}\|/\widetilde{\varepsilon}_B \leqslant \sqrt{2}\|\Delta\|\|\mathbf{c}\|/\varepsilon_{AB}\widetilde{\varepsilon}_B. \quad (5.81)$$

The last inequality is a consequence of the constraints **(SC)** and **(PC)**. Indeed,

$$\|(\mathbb{I}_N - A - B)^{-1}\mathbf{c}\| = \left\| \sum_{i=0}^{\infty} (A+B)^i \mathbf{c} \right\| \leqslant \sum_{i=0}^{\infty} \|(A+B)\|_{op}^i \|\mathbf{c}\| \leqslant \sum_{i=0}^{\infty} (1-\varepsilon_{AB})^i \|\mathbf{c}\| = \frac{\|\mathbf{c}\|}{\varepsilon_{AB}}.$$

Now, by (5.76) and (5.81),

$$\|E\left[(T_A^* H_t - T_A^* H_t^k) \cdot \Delta\right]\| = \|E\left[T_A^* H_{t-k} \cdot vech^*(B^{k\,T} math^*(\Delta))\right]\|$$
$$\leqslant \|T_A^* H_{t-k}\|_{op} \|vech^*\|_{op} \|B\|_{op}^k \|math^*\|_{op} \|\Delta\| \leqslant \frac{2\|\Delta\|\|\mathbf{c}\|}{\varepsilon_{AB}\widetilde{\varepsilon}_B}(1-\widetilde{\varepsilon}_B)^k, \quad (5.82)$$

which proves (3.28). If we want this upper bound for the error to be smaller than a certain $\delta > 0$, we have to make the number of iterations k big enough so that

$$\frac{2\|\mathbf{c}\|}{\varepsilon_{AB}\widetilde{\varepsilon}_B}(1-\widetilde{\varepsilon}_B)^k < \delta \quad \text{that is} \quad (1-\widetilde{\varepsilon}_B)^k = \frac{\delta\varepsilon_{AB}\widetilde{\varepsilon}_B}{2\|\mathbf{c}\|} \leqslant \frac{\delta\varepsilon_{AB}\widetilde{\varepsilon}_B}{2\varepsilon_{\mathbf{c}}}.$$

This relation, together with (5.80) proves the estimate (3.30). ■

Chapter F

Mixtures of GAMs for habitat suitability analysis with overdispersed presence/absence data

with David Pleydell

Abstract

This paper proposes a new approach to species distribution modelling based on unsupervised classification via a finite mixture of GAMs incorporating habitat suitability curves. An tailored EM algorithm is proposed for computing maximum likelihood estimates. Several submodels incorporating various parameter constraints are explored. Simulation studies confirm that under certain constraints, the habitat suitability curves are recovered with good precision. The method is also applied to a set of real data concerning presence/absence of observable small mammal indices collected on the Tibetan plateau. The resulting classification was found to correspond to species-level differences in habitat preference described in previous ecological work.

1 Introduction

Understanding variations in species distribution has remained one of the key challenges in ecology since its conceptualisation as a discipline [**?**,]. It has been natural that ecologists should seek to model species distribution and early models date from the nineteen twenties [**?**,]. Uses of species distribution models (SDMs) in conservation biology include [**?**,]: quantification of environmental niches for species; testing biogeographical, ecological and evolutionary hypotheses; invasive species monitoring; impact assessment for climatic change; prediction of unsurveyed sites for rare species; management support for species reintroduction and recovery; conservation planning; species assemblage modelling; classification of biogeographic or ecogeographic regions; calibration of ecological distance between patches in meta population or gene flow models.

Several techniques have been employed for SDMs including: generalised linear models (GLMs) and their flexible extension generalised additive models (GAMs) ([**?**], [**?**], [**?**]); tree based classification techniques [**?**,]; ordination [**?**,]; eco-niche factor analysis [**?**,]; Bayesian approaches [**?**,]; neural networks [**?**,] and support vector machines [**?**,]. Ecologists have long recognised the bias introduced into SDMs when data are overdispersed with respect to

a simple parametric model such as can arise when strong spatial dependence exists between observations for example ([**?**], [**?**], [**?**]) but the proportion of articles published in ecological journals in which these biases are reasonably corrected for remains low. One problem, particularly in the spatial context, has been the lack of available tools for analysing overdispersed binary or Poisson data. This situation has been slowly changing since the seminal work of [**?**] who introduced the geostatistical concept of Gaussian random fields to the GLM literature to account for spatially smooth sources of overdispersion. Since then appropriate tools have become increasingly more available: the geoRglm library [**?**,] for Bayesian analysis of GLMs with geostatistical priors and the MGCV library for fitting generalised additive mixed models with either geostatistical or spline based random effects using penalised likelihood [**?**,] are just two examples of what is now available for R [**?**,].

A recent review (with online R code) of available techniques for the estimation of Gaussian random fields within a GLM for spatially dependant Bernoulli data [**?**,] suggested that the estimation of spatially structured random effects could be reasonable if the underlying spatial structure was simple relative to the sampling density of the points. However when each curve and bend in a complex hidden surface was sparsely sampled then attempts to estimate the hidden surface proved less successful. The estimation of complicated hidden spatial structure from Bernoulli samples is now recognised to be highly data demanding suggesting that these models might be unreasonable in certain practical situations where logistical constraints limit the quantity of available data. We could ask the question "is it always necessary to estimate continuous spatial random effects plus three or four variogram parameters for binary ecological data sets?" or even "are hidden spatial structures in ecological data sets always smooth?". If the answers to these questions is "no" then perhaps we can simplify and reduce the number of random effects and parameters that we expect to estimate, thereby reducing the demands we place on our datasets. In this paper we attempt to do this using a mixture model approach where the usual single GAM with n continuous random effects might be replaced by say K GAMs. Such a simplification would require a small number of parameters relative to n especially when further constraints between the mixture components are imposed.

Note that here we do not attempt to explicitly model sources of overdispersion. The mixture model approach simply provides a general solution to account for various sources of overdispersion. According to [**?**,] mixture components "correspond to particular zones of support of the true distribution" and thus provide local represention of the likelihood function. While these local supports "do not always posses an individual significance or reality for the particular phenomenon modelled", interpretability can be possible in situations such as discrimination or clustering. This is the case for our model and a real data example is section 5 is found to provide a very natural ecological interpretation.

It is worth noting that the simplification we propose is not necessarily made at the expense of physical interpretation. In a given ecological context a small number of discrete random effects could be a reasonable model for hidden spatial structure or other sources of overdispersion. For example, if the species in question lived in colonies one GAM could represent within colony densities and a second GAM could represent non-colony densities. Similarly, if the observations in question materialised from numerous different processes then a mixture model approach could be expected to outperform it's $K = 1$ counterpart. The most pertinent number of random effects K could then be identified using model selection techniques. Herein lies an additional advantage of our approach, our GAM utilises a simple transformation on covariates and so the parameters for our mixture model can be estimated by maximum likelihood. For highly flexible models such as GAMs with splines or random fields ML is known to be prone to over fitting and penalisations are often imposed to compensate. Since we use a mixture of simple GAMs with relatively limited flexibility we can

use maximum likelihood directly without penalisation. For model comparison statistics such as Akaike Information Criterion (AIC) [**?**,] are therefore readily available.

In the current paper we implement this proposed model simplification in a habitat suitability identification context. Habitat suitability curves are used to identify non-linear species responses along environmental gradients ([**?**],[**?**],[**?**]). The concept is to identify a curve which transforms a continuous environmental variable to a scale more relevant to the distribution of the species in question thereby giving an index of habitat suitability.

2 A generalised additive model for habitat suitability identification

GAMs for habitat suitability detection

Generalised additive models (GAMs) have become popular tools in ecology due to their ability to detect non-linearities. A recent review of GAMs can be found in [**?**,]. The usual approach is to add smooth functions of covariates to the linear predictor of a generalised linear model [**?**,]. We take the simple case,

$$g(\mu_i) = \beta_0 + \beta_1 \mathcal{H}(x_i)$$

where $\mu \equiv E[Y]$, Y follows some distribution of the exponential family, β_0 is the intercept and $\mathcal{H}$ is a smooth function of covariate x. Perhaps the most common choice for $\mathcal{H}$ are spline functions [**?**,] which are highly flexible. Here however we use a much simpler habitat suitability curve to detect a single region within an environmental gradient within which a given species is found in greatest abundance. We avoid the term "niche detection" since we work exclusively in the univariate case in the interest of maintaining simplicity. In our GAM $\mathcal{H}$ is defined as the unimodal transformation

$$\mathcal{H}_{\alpha_1,\alpha_2}(x) = \frac{\left(\frac{x-l}{u-l}\right)^{\alpha_1}\left(\frac{u-x}{u-l}\right)^{\alpha_2}}{\left(\frac{m-l}{u-l}\right)^{\alpha_1}\left(\frac{u-m}{u-l}\right)^{\alpha_2}}.$$

This transformation is a flexible uni-modal mapping from the range $[l, u] \subset \mathbb{R}$ to $[0, 1]$ and is intended to be flexible enough to identify the most pertinent subset of x corresponding to those areas where a species may be found in greatest density. The parameters α_1 and α_2 may take values in $(0, \infty)$ and $\mathcal{H}_{\alpha_1,\alpha_2}(l) = \mathcal{H}_{\alpha_1,\alpha_2}(u) = 0$. The value $m = (u\alpha_1 + l\alpha_2)/(\alpha_1 + \alpha_2)$ locates the mode, i.e. $x = m$ maximises $\mathcal{H}(x)$ such that $\mathcal{H}(x = m) = 1$. As $\{\alpha_1, \alpha_2\} \to (0, 0)$ then $\mathcal{H}_{\alpha_1,\alpha_2}(x) \to 1 \ \forall \ x \in (l, u)$ giving a uniform mapping in the limit. As $\{\alpha_1, \alpha_2\} \to (\infty, \infty)$ then $\int_l^u \mathcal{H}_{\alpha_1,\alpha_2}(x)dx \to 0$.

Transformation (2.1) can be re-parameterised in terms of α_1 (α from here on) and m. This has the advantage over (2.1) of greater orthogonality between parameters plus a more intuitive interpretation of m. The new parameterisation is thus

$$\mathcal{H}_{\alpha,m}(x) = \left(\frac{x-l}{m-l}\right)^{\alpha}\left(\frac{u-x}{u-m}\right)^{\alpha\frac{u-m}{m-l}}.$$

In what follows x represents a continuous index of some environmental gradient such as vegetation biomass, soil moisture, mean daily temperature etc. In practice such an index might be mapped across the study area in raster format.

The mixture of GAMs model

We will now introduce our mixture of GAMs. We will assume that given the vector $(\mathcal{H}(x_1), \ldots, \mathcal{H}(x_n))$ each observation $y_i \in \{1, 0\}$ corresponding to presence/absence, is sampled from the distribution

$$f_{mix}(y_i) = \sum_{k=1}^{K} p_k f_{ik}(y_i),$$

where

$$f_{ik}(y_i) = \pi_{ik}^{y_i}(1 - \pi_{ik})^{1-y_i}$$

and

$$\pi_{ik} = \frac{\exp\left(\beta_{0k} + \beta_{1k}\mathcal{H}_{\alpha_k, m_k}(x_i)\right)}{1 + \exp\left(\beta_{0k} + \beta_{1k}\mathcal{H}_{\alpha_k, m_k}(x_i)\right)}.$$

The unknowns in this model which we will have to estimate for each are $k \in \{1, \ldots, K\}$,

- the probability weights p_k s.t. $\sum_{k=1}^{K} p_k = 1$
- the reals $\beta_{0k} \in \mathbb{R}$ and $\beta_{1k} \in \mathbb{R}^+$
- the parameters (α_k, m_k) of the functions $\mathcal{H}_{\alpha_k, m_k}$ which map $[l, u]$ to $[0, 1]$ and linearise the influence of a bounded continuous index of environmental variation.

The goal of this model is to split the sample into K classes of data with similar statistical properties. It is expected that these classes will reflect to a certain extent the sources of overdispersion within the observed phenomenon at a reasonable computational cost, i.e. without being over demanding of the information available in the data. This formulation is clearly not spatially explicit and so prediction of hidden spatial structure at unsampled locations is not a feature of our model. This is a further step that we will investigate in future work.

3 Estimation and EM algorithm

We now address the question of estimating the unknown parameters of our mixture model. The estimation of the parameters can be obtained using the maximum likelihood approach for which the EM algorithm is well tailored.

Maximum likelihood

We now enter the details of the maximum likelihood approach for estimation in our mixture model. The observed data are couples (y_i, x_i), $i = 1, \ldots, n$. To this sample, we associate a sequence of couples (Y_i, X_i) of independent random variables, $i = 1, \ldots, n$ such that the value of the conditional likelihood taken at $(y_1, \ldots, y_n)$ given the event $\{X_1 = x_1, \ldots, X_n = x_n\}$ may be written as

$$L_{y_1, \ldots, y_n}(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k f_{ik}(y_i),$$

with the f_{ik} given by formula (2.1) and where θ is the vector of unknown parameters, i.e.

$$\theta = (p_1, \ldots, p_K, \beta_{01}, \ldots, \beta_{0K}, \beta_{11}, \ldots, \beta_{1K}, \alpha_1, \ldots, \alpha_K, m_1, \ldots, m_K).$$

The vector of parameters θ can be estimated using the maximum likelihood procedure, i.e.

$$\theta_{ML} \in argmax_{\theta\in\Theta} L_{y_1,\ldots,y_n}(\theta),$$

where Θ is the domain of the likelihood function satisfying

$$\Theta \subset \Big\{\theta = (p_1,\ldots,p_K,\beta_{01},\ldots,\beta_{0K},\beta_{11},\ldots,\beta_{1K},\alpha_1,\ldots,\alpha_K,m_1,\ldots,m_K) \\ \in [0,1]^K \times \mathbb{R}^K \times [0,\infty)^K \times (0,\infty)^K \times (u,l)^K \mid \textstyle\sum_{k=1}^K p_k = 1\Big\}.$$

The domain may also incorporate various additional restrictions on the model such as the possible equalities of certain parameters between classes.

There now remains to notice that a vector θ_{ML} maximizing the conditional likelihood cannot be obtained via a closed form formula. Thus, an iterative algorithm has to be used and in the following section we describe a version of the well known EM algorithm for this purpose.

The EM algorithm

Description of the method

The EM algorithm is a well known conceptual scheme allowing to build recursive procedures that converge towards a set of vectors maximizing the likelihood, or more appropriately here, the conditional likelihood over the domain Θ. EM has been proposed in its present general form by Dempster, Laird and Rubin in [**?**,], hence encompassing several specialised procedures that had been developed in various applications of the maximum likelihood principle. The main reference on EM algorithms, their variants and their applications is the book [**?**,].

The idea underlying the EM algorithm is the following. It is expected that if more information on the observations were available, then optimising the likelihood could be performed easily. The main additional information that we could have in the ecological setting is the class of the mixture to which each observation belongs.

If we denote by Z_i the random index of the mixture component from which observation Y_i was drawn, the so-called complete data is actually given by the triples $(Y_1,Z_1,X_1),\ldots,(Y_n,Z_n,X_n)$. One still has to keep in mind that the Z_i's are actually unobserved and that their only contribution is to provide the right framework underlying the EM procedure. The complete likelihood associated to the complete data is given by

$$L^c_{(Y_1,Z_1),\ldots,(Y_n,Z_n)}(\theta) = \prod_{i=1}^n p_{Z_i}\pi_{iZ_i}^{Y_i}(1-\pi_{iZ_i})^{(1-Y_i)},$$

where π_{iZ_i} is given by (2.1) above. One of the main features of the complete likelihood is that it can usually be optimised in a easier fashion than the plain likelihood. This is the exact reason why statisticians have been using the EM approach.

E Step. Assume that we have a current value of θ, denoted hereafter by $\tilde{\theta}$. Then, one unreachable but tempting goal would be to optimise the complete likelihood. Now here is the crux: the Z_i's are not observed. One sensible way to approximate $\log L^c_{(Y_1,Z_1),\ldots,(Y_n,Z_n)}(\theta)$ then is to take its minimum mean squared error estimator among functions of the Y_i's only. It is well known that the minimum mean squared error estimator is given by the conditional expectation given the Y_i's assuming that the underlying probability is specified by $\tilde{\theta}$, i.e.

$$Q(\theta, \tilde{\theta}) = E_{\tilde{\theta}}\Big[\log L^c_{(Y_1,Z_1),\ldots,(Y_n,Z_n)}(\theta) \mid Y_1 = y_1, \ldots, Y_n = y_n \Big].$$

In the case of our model, this conditional expectation is quite simple to obtain. Indeed, one only needs to know the values of the conditional probabilities for each possible value of Z_i, $i = 1, \ldots, n$ given $Y_1, \ldots, Y_n$ under the model specified by $\tilde{\theta}$. Using Bayes' rule, one obtains

$$P_{\tilde{\theta}}(Z_i = k \mid Y_i = y_i) = \frac{f_{ik}(y_i; \tilde{\theta})\tilde{p}_k}{\sum_{k'=1}^{K} f_{ik'}(y_i; \tilde{\theta})\tilde{p}_{k'}}.$$

Therefore, we obtain that

$$Q(\theta, \tilde{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \Big(\log(p_k) + y_i \log \pi_{ik} + (1 - y_i) \log(1 - \pi_{ik}) \Big) P_{\tilde{\theta}}(Z_i = k \mid Y_i = y_i).$$

M Step. The next step is the choice of the next iterate, θ_{next}. The idea for obtaining a sensible candidate is quite simple: just maximise the approximation of the complete log-likelihood conditionally on the observations $y_1, \ldots, y_n$, i.e.

$$\theta_{next} \in argmax_{\theta \in \Theta} Q(\theta, \tilde{\theta}).$$

Finally the EM algorithm consists of repeating these two steps recursively until the increase of the likelihood obtained between two successive iterates is judged sufficiently small. In the following, we will write the sequence of EM iterates $(\theta^{(\nu)})_{\nu \in \mathbb{N}}$.

Implementation details

Given iterate $\theta^{(\nu)}$ at step ν, the computation of the next iterate is obtained by solving the first order optimality condition

$$\nabla Q(\theta, \theta^{(\nu)}) = 0,$$

where ∇ is the gradient with respect to the vector of variables θ. Cancelling the partial derivatives with respect to the p_k's is easy and gives the same result as in any mixture model of this type, i.e.

$$p_k^{(\nu+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(\nu)}}{\sum_{i}^{n} \sum_{k'=1}^{K} \tau_{ik'}^{(\nu)}} = \frac{1}{n} \sum_{i}^{n} \tau_{ik}^{(\nu)}$$

where $\tau_{ik}^{(\nu)}$ is the posterior probability that $Z_i = k$ given $Y_i = y_i$ under the model parametrised by the current estimate $\tilde{\theta}$. More explicitly,

$$\tau_{ik}^{(\nu)} = P_{\theta^{(\nu)}}(Z_i = k \mid Y_i = y_i) = \frac{f(y_i \mid Z_i = k; \theta^{(\nu)}) p_k^{(\nu)}}{\sum_{k'=1}^{K} f(y_i \mid Z_i = k'; \theta^{(\nu)}) p_{k'}^{(\nu)}}$$

The computation is less straightforward for other components of θ. The gradient of Q with respect to all the other components has been calculated and is given in the Appendix below. The expression for the gradient should convince that no closed form formula can be obtained for the solution of (3.-36). With this respect, our situation is different from the case of Gaussian mixtures for instance where the successive iterations can be computed by hand. Therefore, a computational approach has to be chosen. We used the L-BFGS-B version of function *optim* in the software R to perform this task.

Parameter equality constraints.

The model described thus far is highly flexible. Depending on the application it can be desirable to impose further constraints. Here we consider the modifications required to ensure that certain parameters in θ are constrained to be equal. For example, if $\theta_1, \ldots, \theta_K$ are parameter sets for models $1, \ldots, K$ respectively, a user might impose that $\theta_1, \ldots, \theta_K$ are equivelent with the exception that $\beta_{01} \neq \cdots \neq \beta_{0K}$ which would provide a mixture model representation of a random effects on intercept model. Alternatively, a user might impose that $\theta_1, \ldots, \theta_K$ are equivelent with the exception that $\beta_{11} \neq \cdots \neq \beta_{1K}$ providing what we call a random effects on β_1 model. An example of such a model in a spatial statistics context (i.e with a spatial prior) is known as geographically weighted regression.

To impose such equality constraint on any subset of parameters $S \subset \theta$ it is sufficient to first impose equivalence in the starting values such that $s_e^{(1)} = s^{*(1)} \; \forall \; S_e \in S$ and thereafter ensure that the L-BFGS-B is pressented with a gradient vector in which the derivative of Q w.r.t S_e (see Appendix) is replaced by the mean derivative w.r.t all elements of S, i.e.

$$\frac{\partial Q}{\partial S_e} = \frac{1}{|S|} \sum_{S_{e'} \in S} \frac{\partial Q}{\partial S_{e'}}$$

In what follows we denote as $\theta_{free} \subset \theta : V \in \theta \Rightarrow V \in \theta_{free}$ the subset of free parameters. Put another way, θ_{free} is equivelent to θ without those elements of θ which are redundant under the model constraints, p_K being such a parameter.

We now describe two simulation studies and a real data application of our mode. The first simulation study (4) investigates identifiability under the random effect on intercept parameterisation. The second simulation study (4) investigates the effect of sample size on the precision of parameter estimates under the random effect on β_1 parameterisation. The real data analysis (5) compares four different parameterisations in the analysis of small mammal indices data collected on the Tibetan plateau, Sichuan Province, China.

4 Simulation studies

First study

Description

A dataset was simulated under the following parameters: $K = 2, n = 1000, p_k = \frac{1}{K}, \beta = \{\beta_0, \beta_1\} = \{-4, -2, 2, 2\}, l = -1, u = 1, \alpha = \{10, 10\}, m = \{0.1, 0.1\}$. The covariate x was simulated over a 100×100 pixel raster grid using a zero-mean Gaussian Random Field (GRF) [**?**,], that is, pixel values were drawn from a multivariate Gaussian distribution with covariance between any two pixels $\mathbf{s}_i$ and $\mathbf{s}_j$ defined as a function of the vector $\overrightarrow{\mathbf{s}_i\mathbf{s}_j}$. We used the so called Gaussian covariance function

$$\Sigma_{i,j} = \sigma_0^2 + \sigma_s^2 \exp\left(\frac{-||\mathbf{s}_i - \mathbf{s}_j||_2^2}{a}\right)$$

with nugget (σ_0^2), sill (σ_s^2) and range (a) set to 0, 5 and 15 (pixels) respectively. This simulation was performed in R using the RandomFields library (http://cran.r-project.org/). The realisation of the GRF was subjected to a linear rescaling so that it was bounded by l and u (Fig. F.1 (a)).

To simulate localised clustering of the "hidden" random effect a second GRF was simulated as above but with $a = 5$ (Fig. F.1 (c)). The 50% quantile of this GRF was used to partition the grid into two classes $Z = 1$ and $Z = 2$ (Fig. F.1 (d)). A stratified sampling

was then implemented with n/K sampling locations simulated at random within each of the two classes. An observation y_i was simulated at each location i in the knowledge of the parameters, covariate and z (Fig. F.1 (f)). For the purpose of parameter estimation z_is were assumed unknown and all parameters were constrained to be shared by models 1 and 2 with the exception of the intercepts and mixing probabilities.

The EM algorithm was used to try to re-capture the true parameter values. This was performed using fifty sets of starting values obtained at random under the following rules: $p^{start} \propto Unif(0,1)$; $\beta_0^{start} \sim Unif(-5,5)$;$\beta_1^{start} \sim Unif(0,5)$; $\beta_{01} = \beta_{02}$; $\alpha \sim Unif(10^{-2}, 10^2)$ and $m \sim Unif(l,u)$. In order to maintain the interpretation that $\mathcal{H}$ provides an index of habitat suitability a lower bound of $\beta_1 = 0$ was imposed in the M-step. The EM was run until l_2 norm difference in θ_{free} between successive iterates became lower than a threshold, i.e. $||\theta_{free}^{(v)} - \theta_{free}^{(v-1)}||_2^2 < 10^{-3}$.

Results

Solutions provided by the EM algorithm were clearly clustered in parameter space. This clustering indicates dependency between starting values and the local optima to which the algorithm converges, a characteristic of mixture models that is widely recognised [**?**,]. Table F.1 shows cluster means and variances of parameter estimates and maximised log likelihoods. The optima closest to the true parameter values was optima 3. The HSC mode m was consistently estimated with precision. The algorithm also detected areas of the likelihood which returned more erroneous parameter estimates and yet higher likelihoods than those obtained using the original parameters, i.e. optima 1 and optima 2. In these solutions β_{01} was under estimated and p_1 over estimated. These solutions appear to correspond to degenerate solutions since lowering both the threshold for the stopping rule and the lower bound of β_0 in the L-BFGS-B algorithm resulted in even lower estimates of β_{01} (not shown). The lowest likelihood corresponded to optima 4 where β_{02} becomes over estimated and β_1 underestimated. These solutions arose when the estimates for α became large causing the HSC too narrow thus increasing the proportion of observations for which $\mathcal{H}(x) \approx 0$. The proportion of observations being significantly influenced by variation in the covariate x was thus reduced and β_{02} grew in order to compensate.

Second study

Description

Data was generated according to the method outlined above (section 4) but with $\beta = \{\beta_0, \beta_1\} = \{-2, -2, 0, 2\}$ and $m = \{m_1, m_2\} = \{0.1, 0.4\}$. An image of the resulting η is shown in Fig. **??**. A range of sample sizes was considerred with $n \in \{5000, 4000, 3000, 2000, 1000, 500, 40$ For each sample size n one hundred realisations of Y_n were generated with x fixed. True parameter values were used as starting values and the EM algorithm was used to maximise the likelihood. The EM was stopped after the first iterate within which the square of the l_2 norm of the difference between successive parameter estimates was smaller than a threshold, i.e. $||\theta_{free}^{(v)} - \theta_{free}^{(v-1)}||_2 < 10^{-2}$.

Results

The mean, variance and l_2 norm of the discrepency between true and fitted values are reported in table F.2. In general the fitted values successfully recapture the original parameter values. The largest discrepancies between original and fitted values appear to be for the α parameter which is not surprising since this parameter might realistically take values across

several orders of magnitude. The largest outliers clearly correspond to those estimates derived from the smallest samples where $n = 100$. Otherwise θ is consistently estimated with a satisfactory degree of precision, the precision in $\hat{m}$ being particularly striking.

It is important to note that the l_2 norm of the error tends to zero as sample size increases. So at the same time the proportion of estimators which are consistent to the true parameter values tends to one and the proportion of meaningless estimators such as those encountered in simulation study 1 tends to zero as sample size grows.

5 Small mammal index example

The data

The data analysed here were from transect surveys conducted in the vicinity of Tuanji, a town situated at 4250m altitude on the Tibetan plateau, Shiqu County, Sichuan Province, China. All transects were made in July 2001 and 2002. Investigators walked straight lines and recorded locations of start, stop and turn points with hand held GPS receivers. After each ten pace interval volunteers stopped and recorded presence or absence of holes belonging to *Microtus limnophilus*, *Microtus leucurus*, *Microtus irene* or *Cricetulus kamensis*. Holes of these species are very similar so no attempt was made to identify holes at the species level. A full account of this data can be found in [**?**,]. The aim here was to present a regression analysis of this presence / absence data with respect to the normalised difference vegetation index (NDVI) derived from a Landsat Enhanced Thematic Mapper (ETM) image acquired on 3rd July 2001. The NDVI here is assumed to provide a suitable proxy index for vegetation biomass for the study area and was derived from ETM's red R and infra-red NIR wave bands as follows.

$$NDVI = \frac{NIR - R}{NIR + R}$$

We applied our mixture of HSC GAMs to this data set in the interest of identifying the range of NDVI within which small mammal indices were observed in greatest number. In our analysis we consider the two types of parameter constraints mentioned in section 3. We will refer to these two models as $\mathcal{M}_1$ and $\mathcal{M}_2$ and define these two models as $\mathcal{M}_1 \equiv \{K = 2, \beta_{01} \neq \beta_{02}, \beta_{11} = \beta_{12}, \alpha_1 = \alpha_2, m_1 = m_2\}$ and $\mathcal{M}_2 \equiv \{K = 2, \beta_{01} = \beta_{02}, \beta_{11} \neq \beta_{12}, \alpha_1 = \alpha_2, m_1 = m_2\}$. We also consider two more flexible models defined as $\mathcal{M}_3 \equiv \{K = 2, \beta_{01} = \beta_{02}, \beta_{11} \neq \beta_{12}, \alpha_1 = \alpha_2, m_1 \neq m_2\}$ and $\mathcal{M}_4 \equiv \{K = 2, \beta_{01} = \beta_{02}, \beta_{11} \neq \beta_{12}, \alpha_1 \neq \alpha_2, m_1 \neq m_2\}$.

Results

AIC values and maximum likelihood estimates of parameters under $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and $\mathcal{M}_4$ are presented in Table F.3. $\mathcal{M}_3$ was selected as the most pertinent with respect to these AIC values. Under $\mathcal{M}_3$ two different modes of the HSC were identified. The component with the m-value equal to 0.58 apparently corresponds to areas of greatest biomass since the maximum NDVI within the study area was 0.53. By comparison with $\mathcal{M}_2$ there appears to be evidence of a bimodal response in small mammal indices with respect to the NDVI gradient.

Our results may be better interpreted using Table 4 and Figure 4 in [**?**,] which show trapping frequencies of the four species in various classes of habitat. *Microtus limnophilus* and *Cricetulus kamensis* were the most frequently trapped species. It is clear from Raoul *et al.* that *Microtus limnophilus* and *Microtus leucurus* were more abundant in areas subjected

to lower grazing pressure where vegetation biomass was greater. The two other species, *Cricetulus kamensis* and *Microtus irene* were more abundant in areas of lower vegetation biomass. It can be safely ascertained that the first component in our mixture with $\beta_{11} = 2.96$ and $m_1 = 0.31$ corresponds to indices of *Cricetulus kamensis* and *Microtus irene* whereas the second mixture component with $\beta_{12} = 8.01$ and $m_2 = 0.58$ corresponds to indices of *Microtus limnophilus* and *Microtus leucurus*. Figures F.3 and F.4 map the derived HS indices for these two groups of species.

6 Discussion

This paper proposed a contribution to modelling species distributions using unsupervised classification via a finite mixture of GAMs and an EM algorithm was proposed for deriving maximum likelihood estimates. Several submodels were studied in which the mixture components were assumed to share certain parameter values. Not all of these submodels appeared satisfactorily identifiable. For instance using different intercepts in two components lead to several possible stationary points, moreover, the one with highest likelihood was far from the vector of true values. On the other hand accurate parameter estimates were obtained when the constraint of equal intercepts was imposed as shown in Table F.2.

The method was also applied to a set of real data concerning presence/absence of observable small mammal indices collected on the Tibetan plateau. The AIC was used to determine the best submodel among 4 candidates and the resulting classification was found to confirm trapping results given in Raoul *et al.* about the common response to vegetation biomass of *Microtus limnophilus* and *Microtus leucurus* on the one hand and *Cricetulus kamensis* and *Microtus irene* on the other.

Several improvements are the focus of our current work. In its present state, our model makes a strong scale assumption, that the model assumes that small mammal responses at a point are most pertinently explained using an index calculated from a single pixel. However Lidicker's ROMPA (Ratio of Optimal to Marginal Patch Area) hypothesis [**?**,] describes how population dynamics can change as a function of the proportion of their preferred habitat within a landscape. There lies hidden here a question of scale since, in addition to habitat quality itself, the distribution or abundance of a given species may respond to the spatial arrangement of preferred habitat [**?**,]. The species *Arvicola terrestris* [**?**,], *Microtus arvalis* [**?**,], *Tetrao urogallus* [**?**,] and the cestode *Echinococcus multilocularis* [**?**,] are just some examples of species who's populations appear to respond to landscape level effects. Scale has become an important issue in ecology and [**?**,] reviews it's multifaceted nature. In order to derive a landscape index such as ROMPA the area over which it is to be calculated must be defined. A commonly adopted approach is to calculate the metric in a circular buffer centerred at each observation. There are two problems. First, a suitable buffer size is not always *a priori* apparent. Secondly, the abrupt cutoff and the indicator weighting scheme that such a buffer imposes is most likely an unrealistic representation of reality. With these ecological considerations in mind, a suitable modification of our model might include the additive component

$$\mathcal{H}_k^{\mathcal{B}_i}(x_i) = \frac{1}{\mathcal{N}_{\mathcal{B}_i}} \sum_{j \in \mathcal{B}_i} \omega_{ij} \mathcal{H}_{\alpha_k, m_k}(x_j)$$

where $\mathcal{B}_i$ denotes the subset of pixels falling within a buffer centerred at location i, $\mathcal{N}_{\mathcal{B}_i}$ is its cardinal and weights ω_{ij} are some function of distance. The $\mathcal{H}_k^{\mathcal{B}_i}(x_i)$ terms therefore introduces into the regression equation the spatially weighted mean habitat suitability within an area surrounding each observation. Preliminary experience with this type of enrichment

indicates that the EM algorithm becomes impractically slow as the buffer size increases. Evidently there is a need for faster algorithms than EM and building such improved methods will be the subject of our next efforts.

Finally future work will also be undertaken on the crucial and exciting question of incorporating spatial dependence into our model via using a discrete random field as a prior for Z.

7 Acknowledgements

The research described was supported by Grant Number RO1 TW001565 from the Fogarty International Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Fogarty International Center or the National Institutes of Health.

Small mammal indices data we collected by Patrick Giraudoux, Nadine Bernard, Renaud Scheifler, Dominique Rieffel and Francis Raoul, all of whom are affiliated to the Department of Environmental Biology, University of Franche-Comté, France. The data my be downloaded from `http://www-math.univ-fcomte.fr/pp_Annu/SCHRETIEN/DataEES`.

8 Appendix

In this section, we provide the formulas for the gradient of $Q(\theta, \tilde{\theta})$ in order for the reader to be able to implement our EM algorithm. As described in section 3, the vector θ is composed of the K mixture probabilities p_k's, the K intercepts $\beta_{01}, \ldots, \beta_{0K}$, the K coefficients $\beta_{11}, \ldots, \beta_{1K}$, the shape parameters $\alpha_1, \ldots, \alpha_K$'s and the mode points $m_1, \ldots, m_K$. The derivative with respect to any variable $V_k \in \{\beta_{0k}, \beta_{1k}, \alpha_k, m_k\}$ is given by

$$\frac{\partial Q}{\partial V_k}(\theta, \tilde{\theta}) = \sum_{i=1}^{n} \tau_{ik} \frac{\partial \pi_{ik}}{\partial V_k} \Big(\frac{y_i - \pi_{ik}}{\pi_{ik}(1 - \pi_{ik})} \Big)$$

with

$$\frac{\partial \pi_{ik}}{\partial \beta_{0k}} = \pi_{ik}(1 - \pi_{ik}),$$

$$\frac{\partial \pi_{ik}}{\partial \beta_{1k}} = \pi_{ik}(1 - \pi_{ik}) \mathcal{H}_{\alpha_k, M_k}(x_i),$$

$$\frac{\partial \pi_{ik}}{\partial \alpha_k} = \pi_{ik}(1 - \pi_{ik}) \beta_{1k} \frac{\partial \mathcal{H}_{\alpha_k, m_k}(x_i)}{\partial \alpha_k},$$

$$\frac{\partial \pi_{ik}}{\partial m_k} = \pi_{ik}(1 - \pi_{ik}) \beta_{1k} \frac{\partial \mathcal{H}_{\alpha_k, m_k}(x_i)}{\partial m_k},$$

$$\frac{\partial \mathcal{H}_{\alpha_k, m_k}(x_i)}{\partial \alpha_k} = \mathcal{H}_{\alpha_k, m_k}(x_i) \Big(\log \Big(\frac{x_i - l}{m_k - l} \Big) + \frac{(u - m_k)}{(m_k - l)} \log \Big(\frac{u - x_i}{u - m_k} \Big) \Big)$$

and

$$\frac{\partial \mathcal{H}_{\alpha_k, m_k}(x_i)}{\partial m_k} = \mathcal{H}_{\alpha_k, m_k}(x_i) \alpha_k \frac{(u - l)}{(m_k - l)^2} \log \Big(\frac{u - m_k}{u - x_i} \Big).$$

9 Tables

	$freq.$	$\bar{l}_{y_1,\ldots,y_n}(\theta)$	$\bar{p}_1$	$\bar{\beta}_{01}$	$\bar{\beta}_{02}$	$\bar{\beta}_1$	$\bar{\alpha}$	$\bar{m}$
EM	50							
Optima 1	11	−467.55	0.67	−13.89	−1.21	6.40	20.22	0.11
		(4.5E-3)	(1.1E-3)	(0.91)	(0.010)	(0.20)	(0.48)	(1.1E-4)
Optima 2	1	−468.80	0.93	−15.00	6.32	14.21	1.16	0.11
Optima 3	3	−469.38	0.88	−2.91	−1.49	2.17	8.31	0.10
		(1.4E-3)	(0.021)	(0.065)	(0.026)	(0.024)	(0.026)	(6.5E-5)
Optima 4	3	−507.22	0.89	−2.11	7.55	0.43	186.32	0.76
		(0.047)	(6.2E-3)	(0.051)	(4.90)	(0.36)	(130.00)	(4.7E-3)
Returned NAs	32	-	-	-	-	-	-	-
		(-)	(-)	(-)	(-)	(-)	(-)	(-)
θ_{true}		−472.78	0.5	−4	−2	2	10	0.1

Table F.1: Summarised results for simulation study 1 in which EM was run from fifty sets of starting values. Solutions were clustered in four sets and cluster means and variances of parameter estimates and the observed data log likelihood are reported here. n is the number of times the algorithms stopped close to the reported cluster means. The algorithm returned NA 32 times in 50.

n		$\mid \hat{p}_1 - p_1 \mid$	$\mid \hat{\beta}_0 - \beta_0 \mid$	$\mid \hat{\beta}_{11} - \beta_{11} \mid$	$\mid \hat{\beta}_{12} - \beta_{12} \mid$	$\mid \hat{\alpha} - \alpha \mid$	$\mid \hat{m}_1 - m_1 \mid$	$\mid \hat{m}_2 - m_2 \mid$	$\|\varepsilon\|_2$
5000	μ	6.97e−03	1.04e−01	1.52e−01	2.02e−01	1.19e+00	2.06e−02	1.43e−02	1.37e+00
	σ^2	6.28e−05	8.30e−03	1.42e−02	2.62e−02	1.16e+00	3.34e−04	1.30e−04	9.77e−01
4000	μ	6.99e−03	1.58e−01	1.92e−01	2.37e−01	1.55e+00	2.36e−02	1.48e−02	1.75e+00
	σ^2	4.58e−05	1.81e−02	2.06e−02	3.02e−02	2.12e+00	3.71e−04	1.09e−04	1.80e+00
3000	μ	7.40e−03	1.18e−01	1.88e−01	3.13e−01	1.78e+00	2.68e−02	1.76e−02	1.96e+00
	σ^2	4.51e−05	9.35e−03	2.26e−02	1.63e−01	3.17e+00	7.07e−04	2.25e−04	3.00e+00
2000	μ	1.22e−02	2.18e−01	2.39e−01	4.47e−01	2.15e+00	3.46e−02	1.98e−02	2.38e+00
	σ^2	1.74e−04	2.79e−02	3.77e−02	5.16e−01	3.77e+00	1.45e−03	2.32e−04	3.77e+00
1000	μ	2.03e−02	2.78e−01	3.89e−01	1.16e+00	4.11e+00	6.34e−02	3.53e−02	4.50e+00
	σ^2	5.20e−04	8.13e−02	9.82e−02	5.17e+00	2.09e+01	3.58e−03	8.60e−04	2.46e+01
500	μ	3.80e−02	5.63e−01	1.07e+00	1.87e+00	7.42e+00	7.92e−02	5.70e−02	8.23e+00
	σ^2	2.07e−03	1.21e+00	3.11e+00	8.67e+00	5.69e+01	5.78e−03	2.16e−03	6.22e+01
400	μ	3.89e−02	4.78e−01	9.73e−01	2.13e+00	9.52e+00	8.03e−02	5.85e−02	1.03e+01
	σ^2	2.43e−03	4.99e−01	2.19e+00	1.02e+01	1.31e+02	4.55e−03	4.34e−03	1.35e+02
300	μ	4.94e−02	9.25e−01	1.68e+00	3.22e+00	1.35e+01	1.04e−01	7.85e−02	1.48e+01
	σ^2	2.28e−03	2.52e+00	5.90e+00	1.30e+01	3.24e+02	6.80e−03	8.00e−03	3.20e+02
200	μ	7.41e−02	1.07e+00	3.13e+00	4.52e+00	2.03e+01	8.71e−02	7.76e−02	2.24e+01
	σ^2	4.84e−03	2.95e+00	1.31e+01	1.96e+01	2.83e+02	5.44e−03	5.00e−03	2.61e+02
100	μ	8.47e−02	2.49e+00	5.29e+00	5.80e+00	2.47e+01	9.74e−02	1.04e−01	2.85e+01
	σ^2	6.15e−03	8.60e+00	1.99e+01	1.77e+01	7.99e+02	6.33e−03	5.53e−03	7.11e+02

Table F.2: Results from simulation study two. 100 Monte Carlo simulations were performed at each sample size and parameters were fitted with EM. Absolute errors for each parameter and the l_2 norm of errors in all free parameters were calculated and their MC means and variances are reported here.

model	AIC	l	p	β_0	β_1	α	m
$\mathcal{M}_1$	3565.3	-1776.7	$\{0.30, 0.70\}$	$\{-1.33, -15.00\}$	8.94	56.11	0.4
$\mathcal{M}_2$	3563.7	-1775.8	$\{0.76, 0.24\}$	-2.52	$\{0.00, 13.32\}$	76.45	0.3
$\mathcal{M}_3$	3556.7	-1771.4	$\{0.36, 0.64\}$	-2.68	$\{3.72, 3.87\}$	80.1	$\{0.30, 0$
$\mathcal{M}_4$	3558.6	-1771.3	$\{0.35, 0.65\}$	-2.68	$\{3.76, 3.85\}$	$\{80.10, 80.11\}$	$\{0.30, 0$

Table F.3: Estimates of parameters in θ_{free} under four different sets of constraints. In $\mathcal{M}_1$ β_{02} reached the lower bound used in the L-BFGS-B algorithm, which resembles the behaviour observed in study 1 associated with degenerate solutions. In $\mathcal{M}_2$ the fitted HSC is clearly not degenerate and corresponds to approximately 24% of the observations y. The AIC is reduced when more than one mode is allowed as in $\mathcal{M}_3$ although there is little evidence of the need for relaxing the equality constraint on α as in $\mathcal{M}_4$.

10 Figures

(a) (b)

(c) (d)

(e) (f)

Figure F.1: Dataset generated for simulation study 1. The first GRF (a) was transformed by the HSC to derive habitat suitability (b). The second GRF (c) was split at the 50% quantile to provide and indicator map of where model 1 (blue) and model 2 (red) operate. A map of $g(\mu)$ (e) was derived from (b) and (d) and used to simulate observed data (f).

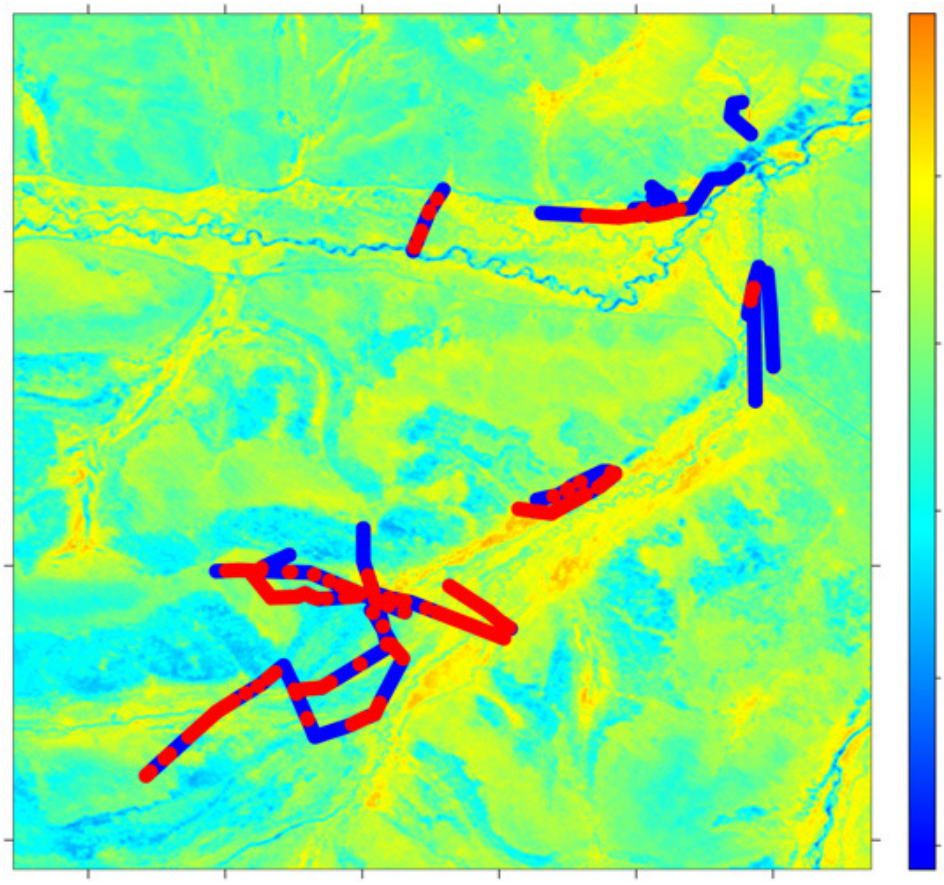

Figure F.2: Normalised difference vegetation index (NDVI) for the Tuanji study area overlayed with transect data on small mammal indices. Red and blue points represent pressence and absence of observable small mammal indeced respectively. Coordinates are in UTM projection.

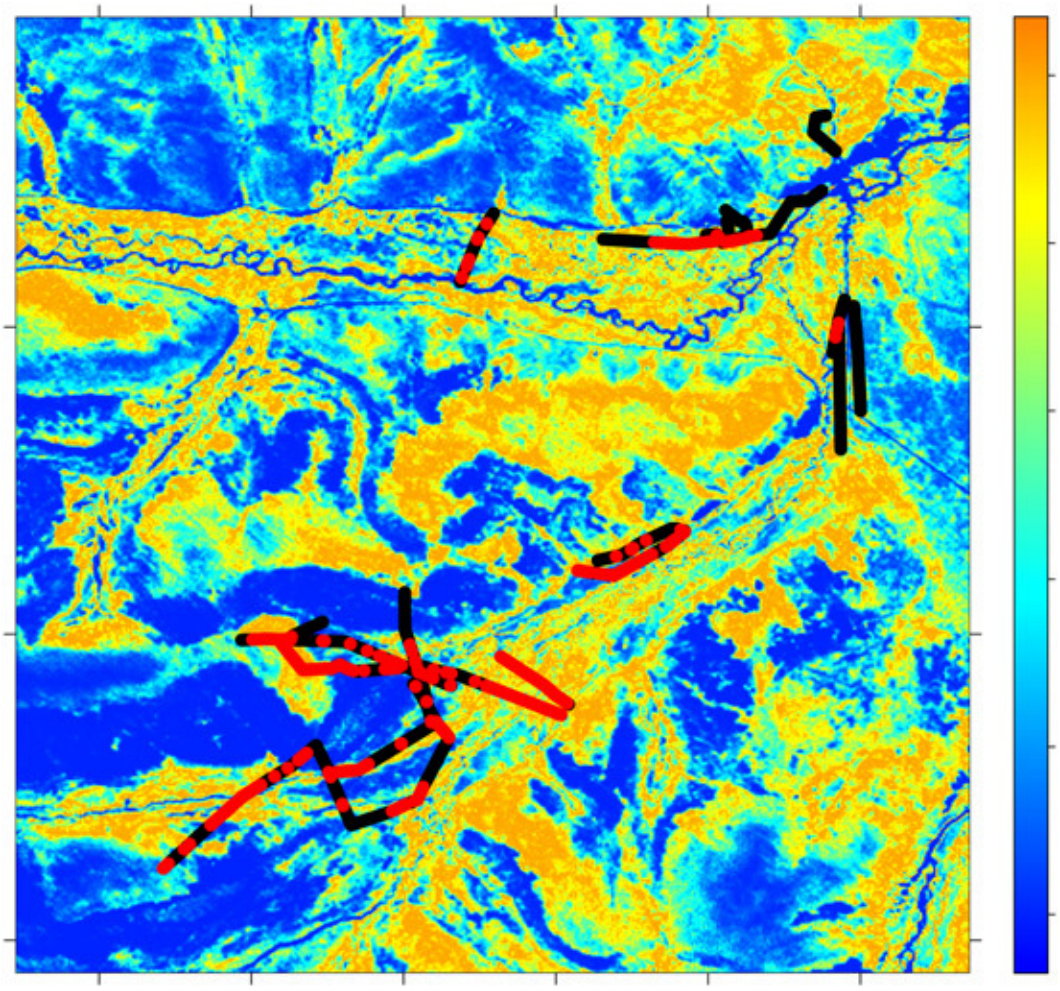

Figure F.3: Habitat Suitability Index derived from the NDVI using ML estimates of $\hat{\alpha}$ and $\hat{m}_1$ from $\mathcal{M}_3$ overlayed with transect data on small mammal indices. Red and black points repressent pressence and absence of observable small mammal indeced respectively.

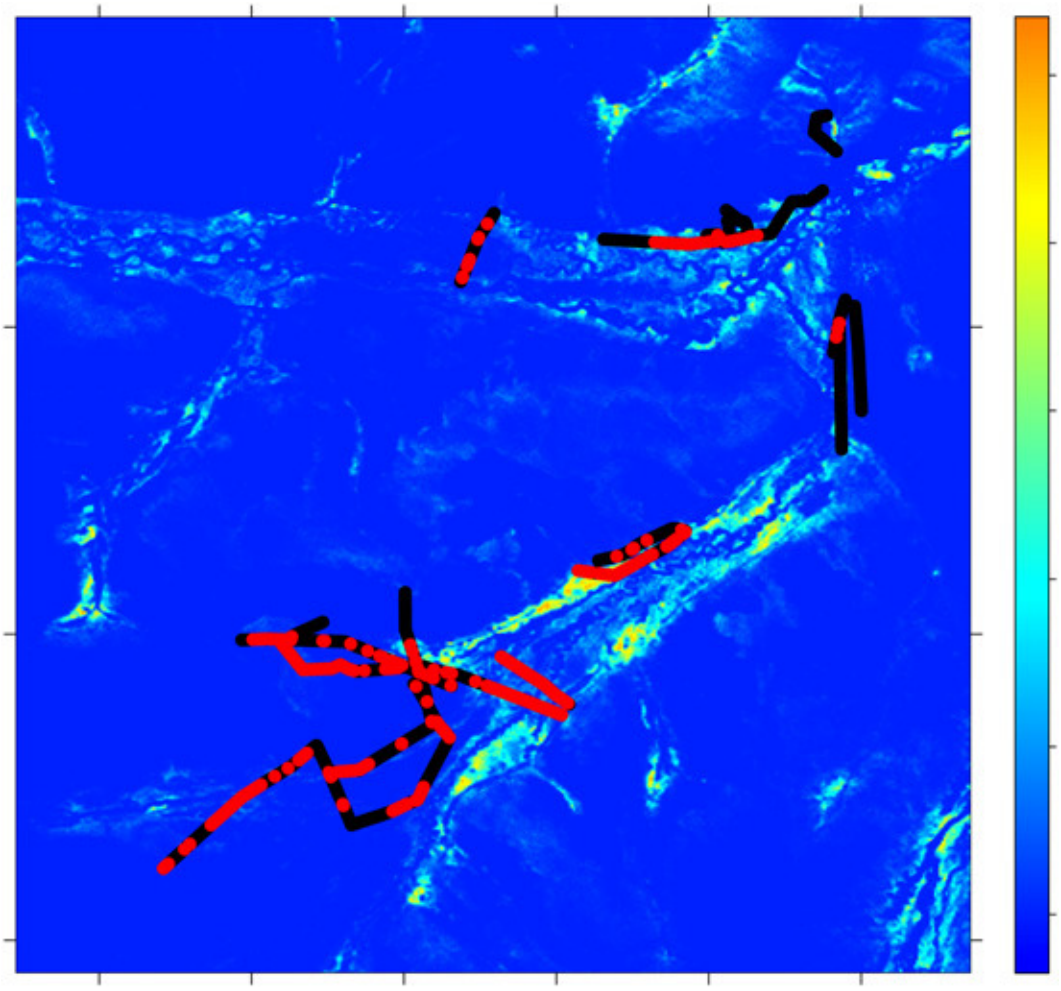

Figure F.4: Habitat Suitability Index derived from the NDVI using ML estimates of $\hat{\alpha}$ and $\hat{m}_2$ from $\mathcal{M}_3$ overlayed with transect data on small mammal indices. Red and black points represent presence and absence of small mammal indices respectively.

Chapter G

An Alternating l1 approach to the compressed sensing problem

Abstract

Compressed sensing is a new methodology for constructing sensors which allow sparse signals to be efficiently recovered using only a small number of observations. The recovery problem can often be stated as the one of finding the solution of an underdetermined system of linear equations with the smallest possible support. The most studied relaxation of this hard combinatorial problem is the l_1-relaxation consisting of searching for solutions with smallest l_1-norm. In this short note, based on the ideas of Lagrangian duality, we introduce an alternating l_1 relaxation for the recovery problem enjoying higher recovery rates in practice than the plain l_1 relaxation and the recent reweighted l_1 method of Candès, Wakin and Boyd.

1 Introduction

Compressed Sensing (CS) is a very recent field of fast growing interest and whose impact on concrete applications in coding and image acquisition is already remarkable. Up to date informations on this new topic may be obtained from the website *http://nuit-blanche.blogspot.com/*. The foundational paper is [1] where the main problem considered was the one of reconstructing a signal from a few frequency measurements. Since then, important contributions to the field have appeared; see [7] for a survey and references therein.

The Compressed Sensing problem

In mathematical terms, the problem can be stated as follows. Let x be a k-sparse vector in $\mathbb{R}^n$, i.e. a vector with no more than k nonzero components. The observations are simply given by

$$y = Ax \tag{1.0}$$

where $A \in \mathbb{R}^{m\times n}$ and m small compared to n with $rankA = m$, and the goal is to recover x exactly from these observations. The main challenges concern the construction of observation matrices A which allow to recover x with k as large as possible for given values of n and m.

The problem of compressed sensing can be solved unambiguously if there is no sparser solution to the linear system (1) than x. Then, recovery is obtained by simply finding the sparsest solution to (1). If for any x in $\mathbb{R}^n$ we denote by $\|x\|_0$ the l_0-norm of x, i.e. the

cardinal of the set of indices of nonzero components of x, the compressed sensing problem is equivalent to

$$\min_{x\in\mathbb{R}^n} \|x\|_0 \quad s.t. \quad Ax = y. \tag{1.0}$$

We denote by $\Delta_0(y)$ the solution of problem (1) and $\Delta_0(y)$ is called a decoder *. Thus, the CS problem may be viewed as a combinatorial optimization problem. Moreover, the following lemma is well known.

Lemma 1.1 *(See for instance [4]) If A is any $m \times n$ matrix and $2k \leqslant m$, then the following properties are equivalent:*

i. The decoder Δ_0 satisfies $\Delta_0(Ax) = x$, for all $x \in \Sigma_k$,

ii. For any set of indices T with $\#T = 2k$, the matrix A_T has rank $2k$ where A_T stands for the submatrix of A composed of the columns indexed by T only.

The l_1 relaxation

The main problem in using the decoder $\Delta_0(y)$ for given observations y is that the optimization problem (1) is NP-hard and cannot reasonably be expected to be solved in polynomial time. In order to overcome this difficulty, the original decoder $\Delta_0(y)$ has to be replaced by simpler ones in terms of computational complexity. Assuming that A is given, two methods have been studied for solving the compressed sensing problem. The first one is the orthognal matching pursuit (OMP) and the second one is the l_1-relaxation. Both methods are not comparable since OMP is a greedy algorithm with sublinear complexity and the l_1-relaxation offers usually better performances in terms of recovery at the price of a computational complexity equivalent to the one of linear programming. More precisely, the l_1 relaxation is given by

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \quad s.t. \quad Ax = y. \tag{1.0}$$

In the following, we will denote by $\Delta_1(y)$ the solution of the l_1-relaxation (1). From the computational viewpoint, this relaxation is of great interest since it can be solved in polynomial time. Indeed, (1) is equivalent to the linear program

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{n} z_i \quad s.t. \quad -z \leqslant x \leqslant z, \quad and \quad Ax = y.$$

The main subsequent problem induced by this choice of relaxation is to obtain easy to verify sufficient conditions on A for the relaxation to be exact, i.e. to produce the sparsest solution to the underdetermined system (1). A nice condition was given by Candes, Romberg and Tao [1] and is called the Restricted Isometry Property. Up to now, this condition could only be proved to hold with great probability in the case where A is a subgaussian random matrix. Several algorithmic approaches have also been recently proposed in order to garantee the exactness of the l_1 relaxation such as in [6] and [5]. The goal of our paper is different. Its aim is to present a new method for solving the CS problem generalizing the original l_1-relaxation of ([1]) and with much better performance in pratice as measured by success rate of recovery versus original sparsity k.

*In the general case where x is not the unique sparsest solution of (1) using this approach for recovery is of course possibly not pertinent. Moreover, in such a case, this problem has several solutions with equal l_0-"norm" and one may rather define $\Delta_0(y)$ as an arbitrary element of the solution set.

2 Lagrangian duality and relaxations

Equivalent formulations of the recovery problem

Recall that the problem of exact reconstruction of sparse signals can be solved using Δ_0 and Lemma 1.1. Let us start by writing down problem (1), to which Δ_0 is the solution map, as the following equivalent problem

$$\max_{z,\, x\in\mathbb{R}^n} e^t z \tag{2.-1}$$

subject to

$$z_i x_i = 0, \quad z_i(z_i - 1) = 0 \quad i = 1, \ldots, n, \text{ and } Ax = y$$

where e denotes the vector of all ones. Here since the sum of the z_i's is maximized, the variable z plays the role of an indicator function for the event that $x_i = 0$. This problem is clearly nonconvex due to the quadratic equality constraints $z_i x_i = 0, \quad i = 1, \ldots, n$.

The standard Semi-Definite Programming (SDP) relaxation scheme

A simple way to construct a SDP relaxation is to homogenize the quadratic forms in the formulation at hand using a binary variable $z_0 = 1$. Indeed, by symmetry, it will suffice to impose $z_0^1 = 1$ since, if the relaxation turns out to be exact and a solution (z_0, z, x) is recovered with $z_0 = -1$, then, as the reader will be able to check at the end of this section, $(-z_0, -z, -x)$ will also solve the relaxed problem. For instance, problem (4.-8) can be expressed as

$$\max_{z,\, x\in\mathbb{R}^n} e^t z z_0 \tag{2.-2}$$

subject to

$$z_i x_i = 0, \quad z_i(z_i - z_0) = 0 \text{ and } z_0 Ax = y$$

for $i = 1, \ldots, n, z_0^2 = 1$.

If we choose to keep explicit all the constraints in problem (2), the Lagrange function can be easily be written as

$$\begin{array}{l} L_{SDP}(w, \lambda, \mu, \nu) = w^t Q w + \sum_{i=1}^n \lambda_i w^t C_i w \\ \quad + \sum_{i=1}^n \mu_i w^t E_i w + v_0 w^t E_0 w \\ \quad + \sum_{j=1}^m \nu_j w^t A_j w - \nu^t y, \end{array}$$

where w is the concatenation of z_0, z, x into one vector, λ (resp. μ and ν) is the vector of Lagrange multipliers associated to the constraints $z_i x_i = 0$, $i = 1, \ldots, n$ (resp. $z_i(z_i - z_0)$, $i = 1, \ldots, n$, and $z_0 a_j^t x = y_j$, $j = 1, \ldots, m$) and where all the matrices Q, A_j, $j = 1, \ldots, m$, E_i, $i = 1, \ldots, n$ and $C_i = 1, \ldots, n$ belong to $\mathbb{S}_{2n+1}$, the set of symmetric $2n+1 \times 2n+1$ real matrices and are defined by

$$Q = \begin{bmatrix} 0 & \frac{1}{2}e^t & 0_{1,n} \\ \frac{1}{2}e & 0_{n,n} & 0_{n,n} \\ 0_{n,1} & 0_{n,n} & 0_{n,n} \end{bmatrix} A_j = \begin{bmatrix} 0 & 0_{1,n} & \frac{1}{2}a_j^t \\ 0_{n,1} & 0_{n,n} & 0_{n,n} \\ \frac{1}{2}a_j & 0_{n,n} & 0_{n,n} \end{bmatrix}$$

for $j = 1, \ldots, m$, where a_i^t is the j^{th} row of A,

$$E_0 = \begin{bmatrix} 1 & 0_{1,n} & 0_{1,n} \\ 0_{n,1} & 0_{n,n} & 0_{n,n} \\ 0_{n,1} & 0_{n,n} & 0_{n,n} \end{bmatrix}, E_i = \begin{bmatrix} 0 & -e_i^t & 0_{1,n} \\ -e_i & 2D(e_i) & 0_{n,n} \\ 0_{n,1} & 0_{n,n} & 0_{n,n} \end{bmatrix}$$

and

$$C_i = \begin{bmatrix} 0 & 0_{1,n} & 0_{1,n} \\ 0_{n,1} & 0_{n,n} & D(e_i) \\ 0_{n,1} & D(e_i) & 0_{n,n} \end{bmatrix}$$

for $i = 1, \ldots, n$ where e_i is the vector with all components equal to zero except the i^{th} which is set to one, e is the vector of all ones, $D(e_i)$ is the diagonal matrix with diagonal vector e_i and where $0_{k,l}$ denotes the $k \times l$ matrix of all zeros. The dual function is given by

$$\theta_{SDP}(\lambda, \mu, \nu) = \sup_{w \in \mathbb{R}^{2n+1}} L(w, \lambda, \mu, \nu),$$

and thus

$$\theta_{SDP}(\lambda, \mu, \nu) = \begin{cases} -\nu^t y \text{ if } Q(\lambda, \mu, \nu) \preceq 0 \\ +\infty \text{ otherwise} \end{cases}$$

with

$$Q(\lambda, \mu, \nu) = w^t Q w + \sum_{i=1}^{n} \lambda_i w^t C_i w + \sum_{i=0}^{n} \mu_i w^t E_i w + \sum_{j=1}^{m} \nu_j w^t A_j w$$

and where $\succeq$ is the Löwner ordering ($A \succeq B$ iff $A - B$ is positive semi-definite). Therefore, the dual problem is given by

$$\inf_{\lambda \in \mathbb{R}^n, \mu \in \mathbb{R}^{n+1}, \nu \in \mathbb{R}^m} \theta_{SDP}(\lambda, \mu, \nu),$$

which is in fact equivalent to the following semi-definite program

$$\inf_{\lambda \in \mathbb{R}^n, \mu \in \mathbb{R}^{n+1}, \nu \in \mathbb{R}^m} -y^t \nu, \tag{2.-11}$$

subject to

$$Q(\lambda, \mu, \nu) \preceq 0. \tag{2.-11}$$

We can also try and formulate the dual of this semi-definite program which is called the bidual of the initial problem. This bidual problem is easily seen after some computations to be given by

$$\max_{X \in \mathbb{S}_{2n+1},\ X \succeq 0} trace(QX) \tag{2.-11}$$

subject to

$$\begin{gathered} trace(A_j X) = y_j,\ j = 1, \ldots, m, \\ trace(E_0 X) = 1, \\ trace(E_i X) = 0 \text{ and } trace(C_i X) = 0,\ i = 1, \ldots, n. \end{gathered} \tag{2.-12}$$

Now, if X^* is an optimal solution with $rank(X^*) = 1$, then

$$X^* = \Big(\pm \begin{bmatrix} z_0^* \\ z^* \\ x^* \end{bmatrix} \Big) \Big(\pm \begin{bmatrix} z_0^* \\ z^* \\ x^* \end{bmatrix} \Big)^t$$

and it can be easily verified that all the constraints in (2) are satisfied. Moreover, we may additionally impose that $z_0^* = 1$ [†]. However, the following proposition ruins the hopes for the occurance of such an agreable situation.

[†]Indeed, if $z_0^* = -1$, multiply by -1 the whole vector $[z_0^*, z^*, x^*]$

Proposition 2.1 *If non empty, the solution set of the bidual problem (2) is not a singleton and it contains matrices with rank equal to $n-m$.*

Proof. Consider the subspace W_0 of $\mathbb{R}^{2n+1}$ as the set of vectors whose $n+1$ first coordinates are equal to zero and such that the last n coordinates form a vector in the kernel of A. Since we assumed that $rankA = m$, we have that $dimW_0 = n - m$. Assume that there exists a solution X^* to (2) with rank less than or equal to $n - m - 1$. Then, it is possible to find a vector w in W_0 with $w^t \perp P_{W_0}(Range(X^*))$. On the other hand, one can easily check that $X^{**} = X^* + ww^t$ satisfies all the bidual constraints and has the same objective value as X^*. Thus, X^{**} is also a solution of the bidual problem and $rankX^{**} = rankX^* + 1$. Iterating the argument up to matrices of dimension equal to $n - 1$, we obtain that the solution set contains matrices with rank equal to $n - m$. To prove non uniqueness of the solution, for any solution matrix X^*, set $X^{***} = X^* + ww^t$ for any choice of w in W_0 and X^{***} is also a solution of the bidual problem. □

Comments on the SDP relaxation

Despite the powerfull Lagrangian methodology behind its construction, the SDP relaxation of the problem has three major drawbacks:

- as implied by Proposition 2.1, the standard SDP relaxation scheme leads to solutions which naturally have rank greater than one which makes it hard to try and recover a nice primal candidate. Moreover, even if the rank problem could be overcome in practice in the case where x is sparse enough, by adding more ad hoc constraints in the SDP, finding the most natural way to do this seemed quite non trivial to us.
- in the case where the SDP has a duality gap, proposing a primal suboptimal solution does not seem to be an easy task.
- the computational cost of solving Semi-Definite Programs is much greater than the cost of solving our naive relaxation, a fact which may be important in real applications.

An utopic relaxation

In order to overcome the drawbacks of the SDP relaxation, we investigate another scheme which may look utopic at first sight. Notice that one interesting variant of formulation (4.-8) could be the following in which the nonconvex complementarity constraints are merged into the unique constraint $\|D(z)x\|_1 = 0$

$$\max_{z\in\{0,1\}^n} e^t z \quad s.t. \|D(z)x\|_1 = 0, \quad Ax = y. \tag{2.-14}$$

Choosing to keep the constraints $Ax = y$ and $z \in \{0,1\}^n$ implicit in (2), the Lagrangian function is given by

$$L(x,z,u) = e^t z - u\|D(z)x\|_1 \tag{2.-14}$$

where $D(z)$ is the diagonal matrix with diagonal vector equal to z. The dual function (with values in $\mathbb{R} \cup +\infty$) is defined by

$$\theta(u) = \max_{z\in\{0,1\}^n,\ Ax=y} L(x,z,u) \tag{2.-14}$$

and the dual problem is

$$\inf_{u\in\mathbb{R}} \theta(u). \tag{2.-14}$$

The main problem with the dual problem (2) is that the solutions to (2) are as difficult to obtain as the solution of the original problem (2) because of the nonconvexity of the Lagrangian function L.

3 The Alternating l_1 method

We now present a generalization of the l_1 relaxation which we call the Alternating l_1 relaxation with better experimental performances than the standard l_1 relaxation and the SDP relaxation.

A practical alternative to the utopic relaxation

Due to the difficulty of computing the dual function θ in the relaxation 2, the interest of this scheme seems at first to be of pure theoretical nature only. In this section, we propose a suboptimal but simple alternating minimization approach.

When we restrict z to the value $z = e$, solving the problem

$$x_(u) = argmax_{z=e,\, x\in\mathbb{R}^n,\, Ax=y} L(x, z, u) \tag{3.-14}$$

gives exactly the solution $\Delta_1(y)$ of the l_1 relaxation. From this remark, and the Lagrangian duality theory above, it may be supected that a better relaxation can be obtained by trying to optimize the Lagrangian even in a suboptimal manner.

Algorithm 2 Alternating l_1 algorithm (Alt-l_1)

Require: $u > 0$ and $L \in \mathbb{N}_*$
  $z_u^{(0)} = e$
  $x_u^{(0)} \in argmax_{x\in\mathbb{R}^n,\, Ax=y} L(x, z^{(0)}, u)$
  $l = 1$
  while $l \leqslant L$ **do**
    $z_u^{(l)} \in argmax_{z\in\{0,1\}^n} L(x_u^{(l)}, z, u)$
    $x_u^{(l)} \in argmax_{x\in\mathbb{R}^n,\, Ax=y} L(x, z_u^{(l)}, u)$
    $l \leftarrow l + 1$
  end while
  Output $z_u^{(L)}$ and $x_u^{(L)}$.

At each step, knowing the value of $z_u^{(l)}$ implies that optimization with respect to $x \in \mathbb{R}^n$ can be equivalently restricted to the set of variables x_i which are indexed by the i's associated with the values of $z_u^{(l)}$ which are equal to one. Thus, the choice of $z_u^{(l)}$ corresponds to adaptive support selection for the signal to recover.

The following lemma states that $z_u^{(l)}$ is in fact the solution of a simple thresholding procedure.

Lemma 3.1 *For all x in $\mathbb{R}^n$, any solution z of*

$$\max_{z\in[0,1]^n} L(x, z, u) \tag{3.-14}$$

satisfies that $z_i = 1$ if $|x_i| < \frac{1}{u}$, 0 if $|x_i| > \frac{1}{u}$ and $z_i \in [0, 1]$ otherwise.

Proof. Problem (3.1) is clearly separable and the solution can be easily computed coordinatewise. □

Open problems

A fully rigorous analysis of the rudimentary Alternating l_1 algorithm for a given u seems quite challenging. However, the two following basic properties hold true:

- Taking $L = 1$ and the suboptimal choice $z_u^{(1)} = e$ gives the standard l_1 relaxation.
- Since the computation of $x_u^{(l)}$ is equivalent to

$$x_u^{(l)} \in argmax_{x \in \mathbb{R}^n,\, Ax=y} \sum_{i \ s.t. (z_u^{(l)})_i = 1} |x_i|, \tag{3.-14}$$

  the number of components of x taken into account in the l_1 objective funtion will hopefully be lower than n.

Based on this, it seems intuitively reasonable to expect that the Alternating l_1 approach should improve over the plain l_1, at least in the case where all the components selected at each iteration have indices in the support of x, just because no useless sparsity penalty is put on the components which are not to be estimated as zero. The simulation experiments below seem to confirm this intuition. Another important question would be to know when does the alternating procedure provide a solution to the optimization problem in the very definition (2) of θ in the case $L = +\infty$, and when this convergence occurs within polynomial time. Based on such results, one could safely try and generalize the approach by associating a Lagrange multiplier to each constraint $|x_i z_i| = 0$ and attack the resulting Lagrangian dual problem using modern non-smooth optimization algorithms such as bundle methods [7].

4 Monte Carlo experiments

Comparison between the success rate of l_1 and Alternating l_1 is shown in Figure 1. Optimization of the Lagrange multiplier u was performed using coarse dichotomic search and we finally used $u = 3$ and $L = 4$ iterations in the Alternating l_1. We also incorporated the results obtained using Boyd, Candes and Wakin's recent proposal called the Reweighted l_1 relaxation. Our proposal outperformed both the plain l_1 and the Reweighted l_1 relaxations for the given data sizes. The programs can be found on the author's webpage at the address *http://stephane.g.chretien.googlepages.com/alternatingl1.*

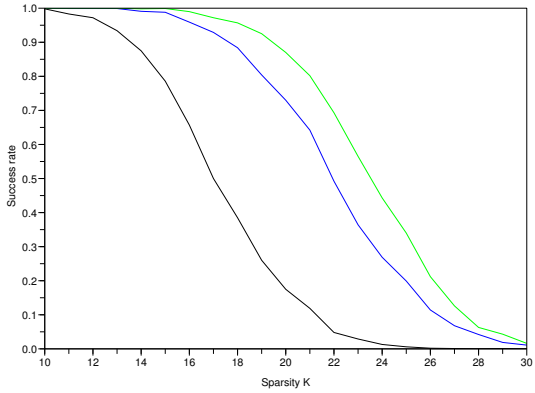

Figure G.1: Rate of success over 1000 Monte Carlo experiments in recovering the support of the signal vs. signal sparsity k for $n = 128$, $m = 50$, $L = 4$, $u = 3$. A and nonnul components of x were drawn from the gaussian $\mathcal{N}(0, 1)$ distribution. The black line is for the l_1 relaxation, the blue line for Boyd, Candes and Wakin's new Reweighted l_1 relaxation with $\varepsilon = .1$, the best value found in [3] and the green line is for our Alternating l_1 relaxation.

Bibliography

[1] Candes, E., Romberg, J. and Tao T., Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Information Theory, 2006, 2, 52, pp. 489–509. (pp. 167 et 168).

[2] Candes, E., Compressive sampling, 2006, 3, International Congress of Mathematics, pp. 1433–1452, EMS. (pp. 167 et 250).

[3] Candes, E., Wakin, M. and Boyd S., Enhancing Sparsity by Reweighted l_1 Minimization, Journal of Fourier Analysis and Applications, 2008, 14, pp. 877–905. (p. 174).

[4] Cohen, A., Dahmen, W. and DeVore R., Compressed sensing and best k-term approximation, J. Amer. Math. Soc. 22 (2009), no. 1, 211–231. (p. 168).

[5] A. d'Aspremont and L. El Ghaoui, Testing the Nullspace Property using Semidefinite Programming, http://arxiv.org/abs/0807.3520. (p. 168).

[6] Juditsky, A. and Nemirovsky, A., On Verifiable Sufficient Conditions for Sparse Signal Recovery via ℓ_1 Minimization, http://arxiv.org/abs/0809.2650. (p. 168).

[7] Hiriart Urruty, J.-B. and Lemaréchal, C., Convex analysis and minimization algorithms II: Advanced theory and bundle methods, Springer- Verlag, 1993, 306, Grundlehren der Mathematischen Wissenschaften.

(p. 173).

Chapter H

Sparse recovery with unknown variance: a LASSO-type approach

with Sébastien Darses.

Abstract

We address the issue of estimating the regression vector β in the generic s-sparse linear model $y = X\beta + z$, with $\beta \in \mathbb{R}^p$, $y \in \mathbb{R}^n$, $z \sim \mathcal{N}(0, \sigma^2 I)$ and $p > n$ when the variance σ^2 is unknown. We study two LASSO-type methods that jointly estimate β and the variance. These estimators are minimizers of the ℓ_1 penalized least-squares functional, where the relaxation parameter is tuned according to two different strategies. In the first strategy, the relaxation parameter is of the order $\widehat{\sigma}\sqrt{\log p}$, where $\widehat{\sigma}^2$ is the empirical variance. In the second strategy, the relaxation parameter is chosen so as to enforce a trade-off between the fidelity and the penalty terms at optimality. For both estimators, our assumptions are similar to the ones proposed by Candès and Plan in Ann. Stat. (2009), for the case where σ^2 is known. We prove that our estimators ensure exact recovery of the support and sign pattern of β with high probability. We present simulations results showing that the first estimator enjoys nearly the same performances in practice as the standard LASSO (known variance case) for a wide range of the signal to noise ratio. Our second estimator is shown to outperform both in terms of false detection, when the signal to noise ratio is low.

1 Introduction

Problem statement

The well-known standard Gaussian linear model reads

$$y \quad = \quad X\beta + z, \tag{1.1}$$

where X denotes a $n \times p$ design matrix, $\beta \in \mathbb{R}^p$ is an unknown parameter and the components of the error z are assumed i.i.d. with normal distribution $\mathcal{N}(0, \sigma^2)$. The present paper aims at studying this model in the case where the number of covariates is greater than the number of observations, $n < p$, and the regression vector β and the variance σ^2 are both unknown.

The estimation of the parameters in this case is of course impossible without further assumptions on the regression vector β. One such assumption is sparsity, i.e. only a few

components of β are different from zero, say s components; β is then said to be s-sparse. There has been a great interest in the study of this problem recently. Recovering the support of β has been extensively studied in the context of Compressed Sensing, a new paradigm for designing observation matrices X. In this framework, it is now a standard fact that matrices X can be found (e.g. with high probability if drawn from sub-Gaussian i.i.d distributions) such that the number of observations needed to recover β exactly is proportional to $s \log(p/n)$.

Existing results in the known variance case

When the variance is known and positive, two popular techniques to estimate the regression vector β are the Least Absolute Shrinkage and Selection Operator (LASSO) [23], and the Dantzig selector [11]. We refer to [3] for a recent simultaneous analysis of these two methods. The standard LASSO estimator $\widehat{\beta}_\lambda$ of β is defined as

$$\widehat{\beta}_\lambda \quad \in \quad \underset{b\in\mathbb{R}^p}{argmin}\ \frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1, \tag{1.2}$$

where $\lambda > 0$ is a regularization parameter controlling the sparsity of the estimated coefficients.

Sparse recovery cannot hold without some geometric assumptions on the dictionary (or the design matrix), as recalled in [24] pp. 4–5. The papers [28] and [22] introduced very pertinent assumptions for the study of variable selection problem using the LASSO in the finite sample (resp. asymptotic) contexts.

One common assumption to study the statistical performance of these estimators is an incoherence property of the matrix X. This means that the coherence of X, i.e. the maximum scalar product of two (normalized) columns of X, is very small. Coherence based conditions appeared first in the context of Basis Pursuit for sparse approximation in [15], [19] and [16]. It then had a significant impact on Compressed Sensing; see [33] and [11].

The recent references [3], [7] and [23] contain interesting assumptions on the coherence in our context of interest, i.e. high dimensional sparse regression. For instance, [3] and [7] require a bound of the order $\sqrt{\log n/n}$ whereas [23] requires a bound of the order $1/s$. The recent paper [14] requires that the coherence of X is less than $Cst/\log p$. Under the additional assumptions that β is sparse and assuming that the support and sign pattern are uniformly distributed, they prove that $\widehat{\beta}$ has the same support and sign pattern as β with probability $1 - p^{-1}((2\pi \log p)^{-1/2} + sp^{-1}) - O(p^{-2\log 2})$. Notice that in [14] the sparsity is not only controlled by the coherence but also by the operator norm $\|X\|$ and implicitly an appropriate choice of the relaxation parameter λ.

Finally, let us notice that the invertibility of the restricted covariance matrix [28] indexed by the signal's true support and the Irrepresentable Condition in [22] can be derived from the incoherence condition in [14]. This would possibly yield suboptimal orders in certain instances though.

Existing results in the unknown variance case

The problem of estimating the variance in the sparse regression model has been addressed in only a few references until now. In [2] the authors analyze in the unknown variance setting AIC, BIC and AMDL based estimators, as well as estimators using a more general complexity penalty. As well known among practitioners, the LASSO procedure, at the price of certain assumptions on X, avoids the enumeration of all subsets of covariates, an

intractable task when the number of covariates is large. This last property motivates the theoretical analysis provided in the present paper.

In [6], a joint estimation procedure for both the regression vector and the variance is proposed. The authors give a detailed study of the risk under quite general conditions. In [35], it is proven in particular that, for the variance estimator of [6], under a compatibility condition introduced in [20], $\lambda\|\beta\|_1/\sigma = o(1)$ if and only if $\widehat{\sigma}/\sigma = (1 + o_{\mathbb{P}}(1))$, for λ such that $\mathbb{P}(\lambda > a\|X^t(Y - X\beta)/n\|_\infty/\sigma) \to 1$ where $a > 1$ is any constant. Moreover, Sun and Zhang [32] study an iterative algorithm, named Scaled LASSO, which is equivalent to the square-root LASSO of Belloni, Chernozhukov and Wang [5], for the joint estimation of the regression coefficients and the noise level σ. In particular, they prove an interesting oracle inequality which shows that the performance of the scaled LASSO method with respect to the risk, is of the same order as for the known-variance case for the usual range of sparsity. However, in these works the problem of support and sign pattern recovery is not addressed.

Our contribution

We study two different strategies in the present paper.

Strategy (A): Plugging in the variance estimator

Our work mainly aims at understanding when the results of [14] extend to the case where σ^2 is unknown. In the case where σ^2 is known, it is proven in [14] that the right order of magnitude for λ is $\sigma\sqrt{\log p}$. We first study the very natural estimator consisting of replacing σ by $\widehat{\sigma} = \|y - X\widehat{\beta}\|_2/\sqrt{n}$ in the expression of λ. As is standard in the study of the LASSO, the regression vector β's coefficients have to be significantly larger than the noise level for exact recovery of the support and sign pattern.

The main differences between the known and the unknown variance cases are summarized in the following table.

Known variance	Unknown variance: Strategy (A)
$\widehat{\beta} \in \underset{b\in\mathbb{R}^p}{argmin}\ \frac{\|y-Xb\|_2^2}{2} + \lambda\|b\|_1$	$\widehat{\beta}_\lambda \in \underset{b\in\mathbb{R}^p}{argmin}\ \frac{\|y-Xb\|_2^2}{2} + \lambda\|b\|_1$
$\lambda = cst\ \sigma\sqrt{\log p}$	Tune λ to $\widehat{\lambda}$ s.t. $: \widehat{\lambda} = C_{var}\widehat{\sigma}\sqrt{\log p}$ $with : \widehat{\sigma}^2 = \frac{\|y-X\widehat{\beta}_{\widehat{\lambda}}\|_2^2}{n}$
Convex problem	*Non convex problem*
Oracle $\widetilde{\beta}$	*Oracle* $(\widetilde{\beta}, \widetilde{\lambda})$
Conditions holding with high probability	*Similar conditions*

Notice that, in this table, $\widehat{\beta}$ is defined via $\widehat{\lambda}$ and $\widehat{\lambda}$ is defined via $\widehat{\beta}$. In other words, $\widehat{\beta}$ and $\widehat{\lambda}$ jointly satisfy a set of optimality conditions. From a numerical viewpoint, $\widehat{\beta}$ and $\widehat{\lambda}$

can be computed iteratively using a simple dichotomic search method described in Section 5.

Strategy (B): Enforcing a trade-off between fidelity and penalty

Another possible strategy can be used to overcome the problem of estimating the regression vector β and the relaxation parameter λ when the variance σ^2 is unknown. This strategy consists of prescribing a trade-off between the fidelity term and the penalty term. More precisely, λ is now an estimator, which is obtained by imposing the constraint $\widehat{\lambda}\|\widehat{\beta}_{\widehat{\lambda}}\|_1/\|y-X\widehat{\beta}_{\widehat{\lambda}}\|_2^2 = C$, where $\lambda \mapsto \beta_\lambda$ is the standard LASSO defined by (1.2). The constant C is selected by the user. Notice that $\widehat{\beta}_{\widehat{\lambda}}$ is sparse (because it is an ℓ_1-penalized least-squares estimator) whereas the least-squares solution is not. This will be confirmed by the simulations results of Section 5.

Enforcing such a trade-off between fidelity and penalty results in a more complex problem from both statistical and computational viewpoints. However, since $\widehat{\lambda}\|\widehat{\beta}_{\widehat{\lambda}}\|_1$ and $\|y-X\widehat{\beta}_{\widehat{\lambda}}\|_2^2$ are, at least approximately, homogeneous functions of σ^2, using such a criterion allows to bypass the estimation of the variance in a first stage. The variance itself could be estimated in a second stage, using the formula $\widehat{\sigma}^2 = \frac{\|y-X\widehat{\beta}_{\widehat{\lambda}}\|_2^2}{n}$.

Known variance	Unknown variance: Strategy (B)
$\widehat{\beta} \in \underset{b\in\mathbb{R}^p}{argmin}\ \frac{\|y-Xb\|_2^2}{2} + \lambda\|b\|_1$	$\widehat{\beta}_\lambda \in \underset{b\in\mathbb{R}^p}{argmin}\ \frac{\|y-Xb\|_2^2}{2} + \lambda\ \|b\|_1$
$\lambda = cst\ \sigma\sqrt{\log p}$	Tune λ to $\widehat{\lambda}$ s.t. : $\widehat{\lambda}\|\widehat{\beta}_{\widehat{\lambda}}\|_1 = C\ \|y - X\widehat{\beta}_{\widehat{\lambda}}\|_2^2$
Convex problem	*Non convex problem*
Oracle $\widetilde{\beta}$	*Oracle* $(\widetilde{\beta}, \widetilde{\lambda})$
Conditions holding with high probability	*Similar conditions + Upper bound on* $\|\beta\|_1$

Results

Our main results are Theorem 2.5, for Strategy (A), and Theorem 2.7, for Strategy (B). Both results can be described as follows. Given an arbitrary $\alpha > 0$, we prove that, for regression vectors β satisfying certain constraints, standard assumptions on the number of observations n and the sparsity s imply that our modified LASSO procedures fail to identify the support and the signs of β with probability at most of the order $p^{-\alpha}$. These results are non-asymptotic and all our constants are explicit.

The coherence assumption on the design matrix made in this paper is readily checkable. Many other currently used assumptions in the literature are based on concentration properties of the extreme singular values of all or most extracted submatrices of X with bounded

number of columns. Yet, some other are based on the concentration of the singular values of the covariance matrix with respect to the covariate's underlying distribution. All such criteria are difficult or impossible to check in practice as opposed to the coherence property.

We neither make any uncheckable assumption on the variance σ^2. The only unverifiable assumptions used in the present work are on the magnitude of the nonzero regression coefficients. As in [14], the set of regressors β which are correctly estimated is constrained by imposing that the magnitude of all nonzero components of β should be greater than the noise level. Moreover, for Strategy (B), our analysis requires the additional assumption that the components of β should not be too large either, the upper bound being in particular a function of C. This result suggests that Strategy (B) is pertinent in low SNR situations only. Simulation experiments at the end of this paper confirm the usefulness of Strategy (B) in the low SNR setting.

The approaches we choose in the present paper have the advantage of allowing a rigourous theoretical investigation of the order of magnitude of maximum admissible sparsity as a function of the problem's dimensions. It would be very interesting to establish similar results based on a LARS-type approach and we leave this open question for further research.

Plan of the paper

The LASSO estimator, the main results Theorem 2.5 and Theorem 2.7, together with the assumptions used throughout the paper are presented in Section 1.0. The proof of Theorem 2.5 is given in Section 3 and the proof of Theorem 2.7 in Section 4. The proofs of certain technical intermediate results are gathered in the Appendix.

Notations

Generalities

When $E \subset \{1, \dots, N\}$, we denote by $|E|$ the cardinal of E. For $I \subset \{1, \dots, p\}$ and x a vector in $\mathbb{R}^p$, we set $x_I = (x_i)_{i \in I} \in \mathbb{R}^{|I|}$. The usual scalar product is denoted by $\langle \cdot, \cdot \rangle$. The notations for the norms on vectors and matrices are also standard: for any vector $x = (x_i) \in \mathbb{R}^N$,

$$\|x\|_2^2 = \sum_{1 \le i \le N} x_i^2 \; ; \quad \|x\|_1 = \sum_{1 \le i \le N} |x_i| \; ; \quad \|x\|_\infty = \sup_{1 \le i \le N} |x_i|.$$

For any matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we denote by A^t its transpose. The set of symmetric real matrices in $\mathbb{R}^{n \times n}$ is denoted by $\mathbb{S}_n$. We denote by $\|A\|$ the operator norm of A. The maximum (resp. minimum) singular value of A is denoted by $\sigma_{\max}(A)$ (resp. $\sigma_{\min}(A)$). Recall that $\sigma_{\max}(A) = \|A\|$ and, if A is invertible, $\sigma_{\min}(A)^{-1} = \|A^{-1}\|$. We use the Loewner ordering on symmetric real matrices: if $A \in \mathbb{S}_n$, $0 \preceq A$ is equivalent to saying that A is positive semi-definite, and $A \preceq B$ stands for $0 \preceq B - A$.

The notations $\mathcal{N}(\mu, \sigma^2)$ (resp. $\chi^2(\nu)$ and $\mathcal{B}(\nu)$) stands for the normal distribution on the real line with mean μ and variance σ^2 (resp. the Chi-square distribution with ν degrees of freedom and the Bernoulli distribution with parameter ν).

Specific notations related to the design matrix X and the estimators

For $I \subset \{1, \dots, p\}$, and a matrix X, we denote by X_I the submatrix whose columns are indexed by I. We denote the range of X_I by V_I and the orthogonal projection onto V_I by $\mathbf{P}_{V_I}$.

The coherence $\mu(X)$ of a matrix X whose columns are unit-norm is defined by

$$\mu(X) = \max_{1\leq i\neq j\leq p} |\langle X_i, X_j\rangle|. \tag{1.2}$$

As in [37], we consider the 'hollow-Gram' matrix H and the selector matrix $R = diag(\delta)$:

$$H = X^tX - I \tag{1.3}$$
$$R = diag(\delta), \tag{1.4}$$

where δ is a vector of length p whose components are i.i.d. random variables following the Bernoulli distribution $\mathcal{B}(s/p)$. In a similar fashion, we define $R_s = diag(\delta^{(s)})$ where $\delta^{(s)}$ is a random vector of length p, uniformly distributed on the set of all vectors with exactly s components equal to 1 and $p-s$ components equal to 0.

The support of $\widehat{\beta}$ is always denoted by $\widehat{T}$.

2 The modified LASSO estimators

In this section, we present the main results on the estimators given by Strategy (A) and Strategy (B), and we discuss the underlying assumptions. Practical computability of these estimators will be studied in Section 5. In particular "tuning λ to $\widehat{\lambda}$" is achieved by finding a zero of a function of λ numerically. We will show in Section 5 that these zero finding problems are computationally very easy to solve.

For any arbitrary value of $\alpha > 0$, Theorem 2.5 (resp. Theorem 2.7), proposes a set of conditions under which exact recovery of the support and sign pattern of β holds with probability at least $1 - O(p^{-\alpha})$ for Strategy (A) (resp. for Strategy (B)).

As will be shortly seen, the magnitude of the nonzero coefficients of β has to satisfy certain constraints: as in [14], one will require for both Strategies that the nonzero components of β are not too small (in fact, slightly above the noise level). In the case of Strategy (B), we will moreover require that the nonzero components of β are not too large. Although this upper bound assumption may seem to argue in disfavor of Strategy (B), computational experiments will later show that this Strategy has much nicer empirical performance when the signal to noise ratio is small. The same computational experiments will also demonstrate that Strategy (A) performs almost as well as a standard LASSO which would know the variance.

Definition of the estimators

To define our estimators, we first need to work with matrices ensuring that the map $\lambda \mapsto \widehat{\beta}_\lambda$, where $\widehat{\beta}_\lambda$ is given by (1.2), is well defined and enjoys special properties, such as continuity.

Definition 2.1 *The matrix X is said to satisfy the Generic Condition if*

$$\left|\langle X_j, X_I(X_I^tX_I)^{-1}\delta_I\rangle\right| < 1, \ \forall \delta \in \{-1,1\}^p, \ \forall I \subset \{1,\dots,p\} \text{ s.t. } X_I \text{ non singular and } \forall j \notin I. \tag{2.5}$$

As from now, we always work under the Generic Condition. We will use the following result about uniqueness of the LASSO estimator.

Proposition 2.2 ([17]) *Assume that X satisfies the Generic Condition. Then, for all $y \in \mathbb{R}^n$, and for all $\lambda \in \mathbb{R}_+$, Problem (1.2) has a unique solution $\widehat{\beta}_\lambda$ and its support $\widehat{T}_\lambda$ is such that $X_{\widehat{T}_\lambda}$ is non singular.*

The following property is proven in Appendix 6:

Lemma 2.3 *Let the Generic Condition hold. Then, almost surely, the map*

$$\left\{ \begin{array}{ccc} (0,+\infty) & \longrightarrow & \mathbb{R}^p \\ \lambda & \longmapsto & \widehat{\beta}_\lambda \end{array} \right.$$

is bounded and continuous. Moreover, $\lambda \longmapsto \|\widehat{\beta}_\lambda\|_1$ *is non-increasing.*

Strategy A

The estimator of strategy A is defined as $\widehat{\beta} := \widehat{\beta}_{\widehat{\lambda}}$ where $\widehat{\lambda}$ verifies the implicit equation

$$\widehat{\lambda}^2 \quad = \quad C_{var} \frac{\left\| y - X\widehat{\beta}_{\widehat{\lambda}} \right\|_2^2}{n} \log p, \tag{2.5}$$

where a relevant range for C_{var} will be given in Theorem 2.5 Eq. (2.5).

The estimators $(\widehat{\beta}, \widehat{\lambda})$ being implicitly defined, it is not clear, at that point, that they exist.

We will see in the sequel that a suitable choice of C_{var} will ensure the existence of the estimators (under the above mentioned assumptions on X).

The uniqueness follows by showing that the map $\Gamma_A : \mathbb{R}_+ \to \mathbb{R}_+$ given by

$$\Gamma_A(\lambda) \quad := \quad \frac{n}{\log p} \, \frac{\lambda^2}{\|y - X_{\widehat{T}_\lambda} \widehat{\beta}_{\widehat{T}_\lambda}\|_2^2},$$

is increasing, which is proven in Appendix 6 .

Strategy A simply reduces to finding the value $\widehat{\lambda}_A$ such that $\Gamma_A(\widehat{\lambda}_A) = C_{var}$. A precise range of interest for C_{var} will be given in Theorem 2.5 below. Moreover, $\widehat{\lambda}$ may be computed using dichotomy search. This scheme is discussed in Section 5.

Remark 2.4 *Recall that in the known variance case, it is often assumed that*

$$\lambda^2 \quad = \quad C_{var} \sigma^2 \log p, \tag{2.5}$$

for some positive constant C_{var}*; see e.g. in [14]. In comparison, Strategy (A) enforces the choice (2.5). This is the empirical analog to (2.5). However, as will appear later in the proof of Theorem 2.5, instead of being an absolute constant,* C_{var} *will have to depend on* n*,* p *and* $\|X\|^2$ *as follows*

$$C_{var} \quad \asymp \quad \frac{n}{p} \|X\|^2.$$

In the case of an i.i.d. Gaussian random design matrix, $\|X\|^2$ *is of the order* p/n *with high probability. Thus* C_{var} *can be basically seen as a constant in the Gaussian setting.*

Strategy B

The estimator of strategy B is defined as $\widehat{\beta} := \widehat{\beta}_{\widehat{\lambda}}$ where $\widehat{\lambda}$ verifies the implicit equation

$$\widehat{\lambda}\|\widehat{\beta}_\lambda\|_1 \quad = \quad C\left\|y - X\widehat{\beta}_{\widehat{\lambda}}\right\|_2^2. \tag{2.5}$$

Again, the estimators $(\widehat{\beta}, \widehat{\lambda})$ are implicitly defined and their existence has to be proven.

Compared to Strategy A, one specificity of Strategy B is that for any value of $C > 0$, existence and uniqueness of the estimators is guaranteed, with no other assumptions than the Generic Condition. Indeed, we show here (cf Lemma 6.5 in the Appendix) that the map Γ_B given by

$$\Gamma_B(\lambda) \quad = \quad \frac{\lambda\|\widehat{\beta}_\lambda\|_1}{\|y - X\widehat{\beta}_\lambda\|_2^2}, \quad \lambda > 0, \tag{2.6}$$

is increasing, continuous and $\Gamma_B((0, +\infty)) = (0, +\infty)$. Thus, there exists a unique value $\widehat{\lambda}_B > 0$ such that $\Gamma_B(\widehat{\lambda}_B) = C$.

Similarly as for Strategy A, the estimation can be performed using a simple dichotomic search described in Section 5.

Main results

Preliminary remarks

The main idea behind the analysis of LASSO-type methods is the following. First, the ℓ_1 penalty promotes sparsity of the estimator $\widehat{\beta}$. Since $\widehat{\beta}$ is sparse, we may restrict the study to the subvector $\widehat{\beta}_{\widehat{T}}$ of $\widehat{\beta}$, resp. the submatrix $X_{\widehat{T}}$ of X, whose components, resp. columns, are indexed by $\widehat{T}$.

Taking this idea a little further, since $\widehat{T}$ is supposed to estimate the true support T of cardinality s, the first kind of result one may ask for is a proof that X_T is far from singular for every possible T. Unfortunately, proving such a strong property with the right order in the upper bound on s, based on incoherence only, seems to be impossible. The idea proposed by Candès and Plan in [14] to overcome this problem is to assume that T is random and then prove that non-singularity occurs with high probability, i.e. for most supports.

Based on this model, the method first consists of proving that X_T satisfies, for $0 < r < 1$,

$$1 - r \leq \sigma_{\min}(X_T) \leq \sigma_{\max}(X_T) \leq 1 + r, \tag{2.7}$$

with high probability. The proof of this property in [14] is based on the Non-Commutative Kahane-Kintchine inequalities. In the present paper, we instead use a result of [13] based on a recent version of the Non-Commutative Chernoff inequality proposed by Tropp [11], in order to obtain better estimates for the involved constants. The most intuitive conditions to prove (2.7) are:

(i) T is a random support with uniform distribution on index sets with cardinal s;

(ii) s is sufficiently small;

(iii) X is sufficiently incoherent.

The main part of the analysis consists of proving that the least-squares oracle estimator, which knows the support ahead of time, satisfies the optimality conditions of the LASSO estimator with high probability. This will prove that the LASSO automatically detects the right support and sign pattern. The proofs of these results highly depend on the quasi-isometry condition (2.7) and similar properties obtained with the same techniques as for (2.7). We also need the sign pattern of β to be uniformly distributed and jointly independent of the support of T. This assumption was already invoked in [14].

Assumptions and main results

As from now, we will work with the following constants:

$$
\begin{aligned}
r \in (0, \frac{1}{2}] & \\
\alpha > 0 \qquad & \text{(controlling the probability that } \|X_T^t X_T - I\| > r) \\
C_{spar} = \frac{r^2}{(1+\alpha)e^2} \qquad & \text{(controlling the sparsity)} \qquad (2.6) \\
C_\mu = \frac{r}{1+\alpha} \qquad & \text{(controlling the coherence)} \qquad (2.7) \\
\kappa = 4\sqrt{1+\alpha} & \qquad (2.7)
\end{aligned}
$$

$$
C_\circ = \ell_\alpha^{-1}\left(10e\frac{1+r}{(1-r)^2}\kappa^2\right) > 0, \qquad \text{(controlling the number of observations)}
$$

where ℓ_α^{-1} is the reciprocal function of the one-to-one continuous map from $(0, +\infty)$ onto $(0, +\infty)$:

$$
\ell_\alpha(x) \quad = \quad xe^{-4\alpha/x}, \quad x > 0.
$$

The constant $C_\circ := C_\circ(\alpha, r)$ is thus well defined (and can be computed numerically by e.g. dichotomy).

The first so-called Coherence condition deals with the minimum angle between the columns of X.

Assumptions 2.1 *(Range and Coherence condition for X) The matrix X has unit ℓ_2-norm columns, is full rank and its coherence verifies*

$$
\mu(X) \quad \leq \quad \frac{C_\mu}{\log p}.
$$

Assumptions 2.2 *(Generic sparse model [14])*

(a) *The support T of β is random and has uniform distribution among all index subsets of $\{1, \ldots, n\}$ with cardinal s,*

(b) *Given T, the sign pattern of β_T is random with uniform distribution over $\{-1, +1\}^s$, and jointly independent of the support.*

Assumptions 2.3 *(Level of sparsity)*

$$s \quad \leq \quad s_0 := \frac{p}{\log p} \, \frac{C_{spar}}{\|X\|^2}.$$

The last condition concerns the magnitude of the nonzero regression coefficients β_j, $j \in T$.

Defining

$$\mathsf{H}_{\alpha,r}^{n,s_0,p} \quad = \quad 4 \frac{\sqrt{n} + \sqrt{2\alpha \log p}}{\sqrt{s_0}} \frac{1-r}{\sqrt{1+r}}. \tag{2.2}$$

we now state the range assumption for the coefficients of β for Strategy (A).

Assumptions 2.4 *(Range condition for β: Strategy (A)) The unknown vector β verifies*

$$\min_{j \in T} |\beta_j| \quad \geq \quad \mathsf{H}_{\alpha,r}^{n,s_0,p} \, \sigma. \tag{2.3}$$

Our main results show that the estimators $\widehat{\beta}$ defined by either Strategy (A) or Strategy (B) recover the support and sign pattern of β exactly with probability of the order $1 - O(p^{-\alpha})$ using similar bounds on the coherence and the sparsity as in [14].

Theorem 2.5 *Set $\alpha > 0$ and $p \geq e^{8/\alpha}$. Let X satisfy the Generic Condition 2.1. Let Assumption 2.2, 2.2, 2.3 and 2.4 hold with*

$$n \quad \geq \quad s \left(C_\circ \log p + 1 \right). \tag{2.4}$$

Then the probability that the estimator $\widehat{\beta}$ defined by Strategy (A) with

$$C_{var} \quad \in \quad \left[\frac{(1-r)^2}{20(1+r)C_{spar}} \, \frac{n}{p} \|X\|^2; \, \frac{(1-r)^2}{2(1+r)C_{spar}} \, \frac{n}{p} \|X\|^2 \right], \tag{2.5}$$

exactly recovers the support and sign pattern of β is greater than $1 - 228/p^\alpha$.

Remark 2.6 *The choice of the constants $1/20$ and $1/2$ in the range of C_{var} is unessential. For application purposes, the practitioner may need to choose a different range of C_{var}, and then decrease or increase these constants. By studying the proof of Theorem 2.5 in Section 3, one notices that e.g. lowering $1/20$ results in lowering the numerical constant 10 in $C_\circ$.*

We now turn to Strategy (B). Let us define for $C > 0$,

$$\mathsf{L}_{\alpha,r,C}^{n,s,p} \quad = \quad \max \left(2 \frac{\sqrt{1+2C}}{C\sqrt{1-r}} \, \frac{\sqrt{n-s} + \sqrt{2\alpha \log p}}{\sqrt{s}} \, , 2 \frac{\sqrt{s} + \sqrt{2\alpha \log p}}{\sqrt{1-r}\,\sqrt{s}} \right) \tag{2.6}$$

$$\mathsf{M}_{\alpha,r,C}^{n,s,p} \quad = \quad \frac{n-s}{\sqrt{\log p}} \, \frac{1}{3\kappa C} \left(\frac{\sqrt{\pi(n-s)}}{p^\alpha} \right)^{\frac{4}{n-s}}. \tag{2.7}$$

The value of C needs to be selected by the user and will be discussed in Section 5.

Let us state the corresponding range assumption for the coefficients of β.

Assumptions 2.5 *(Range condition for β: Strategy (B)) The unknown vector β verifies*

$$\min_{i\in T}|\beta_j| \geq \mathsf{L}^{n,s,p}_{\alpha,r,C}\,\sigma, \tag{2.8}$$

$$\|\beta\|_1 \leq \mathsf{M}^{n,s,p}_{\alpha,r,C}\,\sigma. \tag{2.9}$$

Theorem 2.7 *Set $\alpha > 0$, $p \geq e^{8/\alpha}$ and $c_\circ = \frac{(6\kappa)^2 e}{1-r}$. Choose $C > 0$. Let X satisfy the Generic Condition 2.1. Let Assumptions 2.2, 2.2, 2.3 and 2.5 hold with this value of C and*

$$n \geq c_\circ(1+2C)\; s\log p + s. \tag{2.10}$$

Then the probability that the estimator $\widehat{\beta}$ defined by Strategy (B) exactly recovers the support and sign pattern of β is greater than $1 - 229/p^\alpha$.

The proofs of Theorems 2.5 and 2.7 are forthcoming in Sections 3 and 4.

Important comments

About X

The normalized Gaussian example is instructive. First, when X is obtained from a random matrix with i.i.d. standard Gaussian random entries by normalizing the columns, the coherence is of the order $\sqrt{\log p/n}$ (See below for a short proof). Therefore, taking n of the order of $\log^3 p$ is sufficient for satisfying the Incoherence Assumption 2.2. Second, it is also well known that $\|X\|^2$ is of the order p/n, see e.g. [34]. This suggests in particular that the upper bound (2.6) on the number s of nonzero components of β may be understood in the Gaussian setting as

$$s \leq \frac{p}{\log p}\,\frac{C_{spar}}{\|X\|^2} = O\left(\frac{n}{\log p}\right).$$

Notice that the estimate $\sqrt{\log p/n}$ of the coherence for i.i.d. Gaussian matrices with normalized columns easily follows from the Paul Levy concentration of measure phenomenon on the sphere [25]. Namely, since each normalized column is Haar distributed on the unit sphere, one has

$$\mathbb{P}\left(|\langle X_j, X_{j'}\rangle| \geq u\right) = \mathbb{E}\,\mathbb{P}\left(|\langle X_j, X_{j'}\rangle| \geq u \mid X_{j'}\right) \leq 2\exp\left(-cn\,u^2\right),$$

for some constant $c > 0$. The union bound then gives

$$\begin{aligned}\mathbb{P}\left(\max_{1\leq j<j'\leq p}|\langle X_j, X_{j'}\rangle| \geq u\right) &\leq \frac{p(p-1)}{2}\cdot 2\exp\left(-cn\,u^2\right),\\ &\leq \exp\left(-cn\,u^2 + 2\log p\right).\end{aligned}$$

Hence, this last quantity is less that $p^{-\alpha}$ for $u \geq \sqrt{\log p/n}\sqrt{(\alpha+2)/c}$.

An interesting question concerns the pertinence of the coherence for the problem of variable selection using the LASSO. The work of [28] shows through numerical investigations that certain conditions on the matrix X (requiring in particular the knowledge of the true signal's support, without any statistical assumptions on beta though), allow to deduce sharp

bounds on the minimum sample size needed for exact support recovery. When the true support is not known ahead of time, conditions such as the ones in [28] are required to hold uniformly or at least for most support with high probability. Proving such a property for matrices more general than i.i.d. Gaussian matrices implies loosing sharp bounds on the minimum sample size. The advantage of the coherence over such assumptions is that it can be computed very easily for any given matrix. The main drawback is that the resulting bounds on the minimum sample size might not be sharp.

One may also consult Section 3 in [1] where the authors provide some interesting examples of matrices (frames), selected at random from various libraries, enjoying small coherence (called in their paper worst case coherence).

Order of $\mathsf{H}_{\alpha,r}^{n,s_0,p}$

In the case where X is i.i.d. Gaussian, the order of s_0 is $n/\log p$ and thus the order of $\mathsf{H}_{\alpha,r}^{n,s_0,p}$ is $\sqrt{\log p}$, just as in [14]. Indeed,

$$\mathsf{H}_{\alpha,r}^{n,s_0,p} \asymp \frac{\sqrt{n}+\sqrt{2\alpha\log p}}{\sqrt{\frac{n}{\log p}}} \asymp \sqrt{\log p}.$$

About C and $\mathsf{L}_{\alpha,r,C}^{n,s,p}$

Increasing the upper bound (2.9) on the magnitude of the β_j's via decreasing the constant C also results in increasing the lower bound (2.8). Therefore, C governs a sliding window inside which the coefficients of β can be recovered by the LASSO. Moreover for a given n, one can decrease the lower bound $\mathsf{L}_{\alpha,r,C}^{n,s,p}$ in Eq. (2.6) by increasing C. This would result on a smaller sparsity in Eq. (2.26). Taking C as $C \sim n/(s\log p)$ implies the usual order $\sqrt{\log p}$ for the minimum of beta's (See Eq. (2.6)). If one wants to specify C in a way that is independent of s one may run the risk of prescribing an incorrect order for $\mathsf{L}_{\alpha,r,C}^{n,s,p}$ as a function of n. This technical issue should however be considered as of theoretical interest only and not so much of a problem in practice. As an analogy, consider the plain LASSO with known variance: there exists a universal way of choosing the parameter λ, but many practitioners use the LARS instead in order to explore all the supports occuring on the λ-trajectory and compare them using a standard model selection procedure (AIC, BIC, Foster and George, etc). In the same manner, one could also vary the value of C and compare all supports on this trajectory. In this spirit, our simulation experiments show the histogram of recovered and incorrectly detected components over a large range of values of C. One nice surprise is that Strategy (B) is quite robust vs. the actual choice of C at such a low signal to noise ratio level.

About the constants C_{spar} and C_μ

Let us compare the numerical values of these constants to the one obtained in [14].

One of the various constraints on the rate α in [14] is given by the theorem of Tropp in [37]. In this setting,

$$\begin{aligned} \alpha &= 2\log 2 \\ r &= 1/2, \end{aligned}$$

the author's choice of $1/2$ being unessential. To obtain such a rate α, they need to impose the r.h.s. of (3.15) in [14] to be less than $1/4$, that is:

$$30C_\mu + 13\sqrt{2C_{spar}} \quad \leq \quad \frac{1}{4}. \tag{2.4}$$

In particular, $13\sqrt{2C_{spar}} \leq \frac{1}{4}$, so $C_{spar} < 1.19 \times 10^{-4}$. Let us choose C_{spar} close to this maximum allowed, say $1.18\ 10^{-4}$. The corresponding greatest value of C_μ is then $\frac{1}{4} - 13\sqrt{2 * 1.18\ 10^{-4}}$:

$$C_{spar} \simeq 1.18\ 10^{-4}, \quad C_\mu \simeq 1.7\ 10^{-3}.$$

(The additional condition coming from the end of the proof of [14, Lemma 3.5], that is $\frac{3}{64C_\mu^2} = 2\log(2)$, is not limiting since $\sqrt{3/(128\log 2)} \gg 1.7\ 10^{-3}$.)

Our theorem allows to choose any rate $\alpha > 0$. To make a fair comparison, let us also choose $\alpha = 1.5 > 2\log 2$ and $r = 1/2$. We obtain:

$$C_{spar} \simeq 1.4\ 10^{-2}, \quad C_\mu = 0.2.$$

About the Generic Sparse Model

The Generic Sparse Model was proposed in [14] for a precise analysis of the support recovery properties of the LASSO estimator. It is natural to use the same model in the present study which extends the LASSO to Strategies (A) and (B) which incorporate the estimation of λ when the variance is unknown. The sparsity assumption is quite natural in several applications ranging from gene expression analysis to image reconstruction and inverse problems. However, the assumption that the signs of the components of β are independent is a bit harder to justify in practice. We use this assumption in the sequel only for the sake of simplifying the mathematical analysis. Further work should be devoted to relaxing this assumption in the future. Obviously, assumming some kind of randomness in the construction of X could help for this purpose.

3 Proof of Theorem 2.5

The proof is divided into several steps. The main two steps are as follows. First, we provide the description and consequences of the optimality conditions for the standard LASSO estimator as a function of λ. Second, we prove that these optimality conditions are satisfied by a simple and natural oracle estimator.

Enforcing the invertibility assumption

We recall the basic result we proved in [13] regarding the invertibility of random submatrices via the Non-commutative Chernoff Inequality.

Theorem 3.1 *Let $r \in (0,1)$, $\alpha \geq 1$. Let X be a full-rank $n \times p$ matrix and s be positive integer, such that*

$$\begin{aligned} \mu(X) \quad &\leq \quad \frac{r}{2(1+\alpha)\log p} \\ s \quad &\leq \quad \frac{r^2}{4(1+\alpha)e^2} \ \frac{p}{\|X\|^2 \log p}. \end{aligned}$$

Let $T \subset \{1, \dots, p\}$ be a set with cardinality s, chosen randomly from the uniform distribution. Then the following bound holds:

$$\mathbb{P}\left(\|X_T^t X_T - I_s\| \geq r\right) \quad \leq \quad \frac{216}{p^\alpha}. \tag{3.1}$$

By Theorem 2.1, we have

$$(1+r)^{-1} \quad \leq \quad \|\left(X_T^t X_T\right)^{-1}\| \quad \leq \quad (1-r)^{-1} \tag{3.2}$$

$$(1-r)^{1/2} \quad \leq \quad \|X_T\| \quad \leq \quad (1+r)^{1/2} \tag{3.3}$$

with probability greater than $1 - 216\ p^{-\alpha}$. Thus, throughout this section, we will assume that (3.2) and (3.3) hold, i.e. we will reduce all events considered to their intersection with the event that (3.2) and (3.3) are satisfied.

The oracle estimator for $\widehat{\beta}$ and $\widehat{\lambda}$

We now discuss the next step of the proof of Theorem 2.5, which consists of studying some sort of oracle estimators for β which enjoys the property of knowing the support T of β ahead of time.

For a given $\widetilde{\lambda}$, one might like to consider the following oracle for $\widehat{\beta}$:

$$\overline{\beta} \quad \in \quad \underset{b \in \mathcal{B}}{argmax} \ -\frac{1}{2}\|y - Xb\|_2^2 - \widetilde{\lambda}\|b\|_1, \tag{3.4}$$

where

$$\mathcal{B} \quad = \quad \{b \in \mathbb{R}^p,\ supp(b) = T,\ sign(b) = sign(\beta_T)\}.$$

However, it is not so easy to derive a closed form expression for $\overline{\beta}$. Therefore, it might be more interesting to consider instead the following oracle:

$$\widetilde{\beta} \quad \in \quad \underset{b \in \mathbb{R}^p,\ supp(b)=T}{argmax} \ -\frac{1}{2}\|y - Xb\|_2^2 - \widetilde{\lambda}\ sign(\beta_T)^t b. \tag{3.4}$$

Indeed, $\widetilde{\beta}$ satisfies

$$X_T^t\left(y - X_T\widetilde{\beta}_T\right) - \widetilde{\lambda}\ sign(\beta_T) \quad = \quad 0,$$

and we obtain that $\widetilde{\beta}$ is given by

$$\widetilde{\beta}_T \quad = \quad \left(X_T^t X_T\right)^{-1}\left(X_T^t y - \widetilde{\lambda}\ sign(\beta_T)\right). \tag{3.4}$$

This formula is the same as in the proof of Th. 1.3 in [14], but here, $\widetilde{\lambda}$ is a variable.

Now let us recall that in the known variance case, Candès and Plan assume that

$$\lambda^2 \quad = \quad C_{var}\sigma^2 \log p, \tag{3.5}$$

for some positive constant C_{var}. It is then relevant to seek our oracle $\widetilde{\lambda}$ as:

$$\widetilde{\lambda}^2 \quad = \quad C_{var}\ \frac{\|y - X_T\widetilde{\beta}_T\|_2^2}{n}\ \log p. \tag{3.6}$$

Replacing $\widetilde{\beta}$ by its value (3.4), we obtain

$$C_{var} \left\| y - X_T \left(X_T^t X_T\right)^{-1} \left(X_T^t y - \widetilde{\lambda}\ sign(\beta_T)\right)\right\|_2^2 \quad = \quad \frac{n}{\log p}\ \widetilde{\lambda}^2.$$

Thus,

$$C_{var}\ \|\mathbf{P}_{V_T^\perp} y + \widetilde{\lambda} X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T)\|_2^2 \quad = \quad \frac{n}{\log p}\ \widetilde{\lambda}^2,$$

and using the orthogonality relations, we obtain

$$C_{var}\ \|\mathbf{P}_{V_T^\perp} y\|_2^2 + \widetilde{\lambda}^2 C_{var}\ \|X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T)\|_2^2 \quad = \quad \frac{n}{\log p}\ \widetilde{\lambda}^2,$$

which is equivalent to

$$\widetilde{\lambda}^2 \quad = \quad \frac{\|\mathbf{P}_{V_T^\perp} z\|_2^2}{\frac{n}{C_{var}\log p} - \ \|X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T)\|_2^2} \tag{3.4}$$

We henceforth work with this definition of $\widetilde{\lambda}$. Notice that $\widetilde{\lambda}$ is well defined whenever

$$C_{var} \quad \leq \quad \frac{n}{\|X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T)\|_2^2\ \log p}. \tag{3.5}$$

The choice of C_{var} will be done in the next section.

Study of the oracle $\widetilde{\lambda}$

In this section, we provide a confidence interval for $\widetilde{\lambda}$. In particular, the first subsection shows that $\widetilde{\lambda}$ is well defined.

Bounds on $\|X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T)\|_2^2$

Using the lower bound on $\sigma_{\min}(X_T)$ and the upper bound on $\sigma_{\max}(X_T)$ given by (3.2) and (3.3), we have, with high probability:

$$\frac{1-r}{(1+r)^2}\ s \leq \quad \|X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T)\|_2^2 \quad \leq \frac{1+r}{(1-r)^2}\ s. \tag{3.6}$$

We write the choice of C_{var} made in (2.5) as

$$\frac{1}{20}\frac{(1-r)^2}{1+r}\frac{n}{s_0\ \log p} \ \leq\ C_{var} \ \leq\ \frac{1}{2}\frac{(1-r)^2}{1+r}\frac{n}{s_0\ \log p}, \tag{3.7}$$

where s_0 is the maximum sparsity allowed in Inequality (5.-22), namely,

$$s_0 \quad = \quad \frac{p}{\log p}\ \frac{C_{spar}}{\|X\|^2}.$$

In particular, the condition (3.5) is satisfied which guarantees that $\widetilde{\lambda}$ is indeed well defined.

Bounds on $\|\mathbf{P}_{V_T^\perp} z\|_2$

Using some well known properties of the χ^2 distribution recalled in Lemma 6.2 in the Appendix, we obtain that

$$\mathbb{P}\left(\|\mathbf{P}_{V_T^\perp}(z)\|_2/\sigma \geq \sqrt{n-s}+\sqrt{2t}\right) \quad \leq \quad \exp(-t) \tag{3.7}$$

and

$$\mathbb{P}\left(\|\mathbf{P}_{V_T^\perp}(z)\|_2^2/\sigma^2 \leq u(n-s)\right) \quad \leq \quad \frac{2}{\sqrt{\pi(n-s)}}\,(u\ e/2)^{\frac{n-s}{4}}\,. \tag{3.8}$$

Tune u such that the r.h.s. of (3.5) equals $2/p^\alpha$, i.e.

$$u \quad = \quad \frac{2}{e}\left(\frac{\sqrt{\pi(n-s)}}{p^\alpha}\right)^{4/(n-s)}\,.$$

Thus, we obtain that

$$\frac{\|\mathbf{P}_{V_T^\perp} z\|_2^2}{\sigma^2} \leq \left(\sqrt{n-s}+\sqrt{2\log(\frac{p^\alpha}{2})}\right)^2 \leq \left(\sqrt{n-s}+\sqrt{2\alpha\log p}\right)^2 \tag{3.8}$$

and

$$\frac{\|\mathbf{P}_{V_T^\perp} z\|_2^2}{\sigma^2} \quad \geq \quad \frac{2(n-s)}{e}\left(\frac{\sqrt{\pi(n-s)}}{p^\alpha}\right)^{4/(n-s)} \tag{3.9}$$

with probability greater than or equal to $1-2p^{-\alpha}$.

Bounds on $\tilde{\lambda}$

Lemma 3.2 *The following bounds hold:*

$$\widetilde{\lambda} \quad \leq \quad \sigma\ \frac{1-r}{\sqrt{1+r}}\ \frac{\sqrt{n-s}+\sqrt{2\alpha\log p}}{\sqrt{s_0}} \tag{3.10}$$

$$\widetilde{\lambda} \quad \geq \quad \kappa\ \sigma\ \sqrt{\log p}. \tag{3.11}$$

Proof. Recall that $0 \leq s \leq s_0$. From (3.7), we have

$$C_{var} \quad \leq \quad \frac{1}{2}\frac{(1-r)^2}{1+r}\frac{n}{s_0\ \log p}.$$

We then obtain, by virtue of (3.4) and the upper bound in (3.6),

$$\begin{aligned}\widetilde{\lambda}^2 \quad &\leq \quad \frac{\|\mathbf{P}_{V_T^\perp} z\|_2^2}{2s_0\frac{1+r}{(1-r)^2} - \|X_T\,(X_T^t X_T)^{-1}\,sign(\beta_T)\|_2^2} \\ &\leq \quad \frac{\|\mathbf{P}_{V_T^\perp} z\|_2^2}{2s_0\frac{1+r}{(1-r)^2} - s_0\frac{1+r}{(1-r)^2}}.\end{aligned}$$

Using the bound (3.8), we deduce (3.10).

On the other hand, the bound (3.9) and

$$\frac{n}{C_{var}\log p} \leq 20\frac{1+r}{(1-r)^2}s_0,$$

yield

$$\widetilde{\lambda}^2 \geq \frac{2(n-s)}{e}\left(\frac{\pi(n-s)}{p^{2\alpha}}\right)^{2/(n-s)}\frac{\sigma^2}{20\frac{1+r}{(1-r)^2}s_0}.$$

From (2.4), we know that n verifies

$$\frac{n-s}{s_0} \geq \frac{n-s_0}{s_0} \geq C_\circ \log p. \tag{3.7}$$

Thus, noting that $(\pi(n-s))^{2/(n-s)} \geq 1$,

$$\widetilde{\lambda}^2 \geq \frac{(1-r)^2}{10e(1+r)}p^{-4\alpha/(n-s)}C_\circ\sigma^2\log p.$$

Writing $p^{-4\alpha/(n-s)} = e^{-4\alpha\log p/(n-s)}$ and, using (3.7) again,

$$\frac{\log p}{n-s} \leq \frac{\log p}{n-s_0} \leq \frac{1}{s_0\ C_\circ} \leq \frac{1}{C_\circ},$$

we obtain

$$p^{-4\alpha/(n-s)} \geq e^{-4\alpha/C_\circ}.$$

Therefore,

$$\widetilde{\lambda}^2 \geq \frac{(1-r)^2}{10e(1+r)}e^{-4\alpha/C_\circ}C_\circ\sigma^2\log p. \tag{3.5}$$

Let us recall that the constant $C_\circ$ has been precisely chosen to satisfy

$$\ell_\alpha(C_\circ) = C_\circ e^{-4\alpha/C_\circ} = 10e\ \frac{1+r}{(1-r)^2}\kappa^2.$$

As a conclusion, we have just proved (3.11). □

Candès and Plan's conditions

To obtain the exact recovery of the support and sign patterns of β, we will need similar bounds as the ones in [14, Section 3.5]. Namely,

(i) $\|(X_T^tX_T)^{-1}X_T^tz\|_\infty \leq \kappa\ \sigma\sqrt{\log p}$

(ii) $\|(X_T^tX_T)^{-1}sign(\beta_T)\|_\infty \leq 3$

(iii) $\|X_{T^c}^tX_T(X_T^tX_T)^{-1}sign(\beta_T)\|_\infty \leq \frac{1}{4}$

(iv) $\|X_{T^c}^t\left(I - X_T(X_T^tX_T)^{-1}X_T^t\right)z\|_\infty \leq \kappa\ \sigma\sqrt{\log p}$

(v) $\|X_T^tX_T - I_s\| \leq r.$

When $r = \frac{1}{2}$, these conditions were proven to hold with high probability in [14] based on previous results due to Tropp [37]. Most of the proofs that these conditions hold with high probability are the same as in [14] up to some slight improvements of the constants.

Proposition 3.3 *The bounds (i-iv) hold with probability at least* $1 - 10/p^{\alpha}$. *Condition (v) holds with probability at least* $1 - 216/p^{\alpha}$.

extbfProof. See Section 6 in the Appendix. □

Last step of the proof

We now conclude the proof using the strategy announced in the beginning of this section:

(i) We prove that the proxies $\widetilde{\beta}$ and $\widetilde{\lambda}$ satisfy the optimality conditions (6.-35) and (6.-34), from which we deduce that $\widehat{\beta} = \widetilde{\beta}$ and $\widehat{\lambda} = \widetilde{\lambda}$.

(ii) Since the proxy $\widetilde{\beta}$ has the right support and sign patterns, we conclude that $\widehat{\beta}$ exactly recovers these features as well.

$\widetilde{\beta}$ and β have the same support and sign pattern

First, it is clear that $\widetilde{\beta}$ and β have the same support. Next, we must prove that $\widetilde{\beta}$ has the same sign pattern as β. Use Proposition 3.3 to obtain

$$\begin{aligned} \|\widetilde{\beta}_T - \beta_T\|_\infty &\leq \|(X_T^t X_T)^{-1} X_T^t z\|_\infty + \widetilde{\lambda}\, \|(X_T^t X_T)^{-1} sign(\beta_T)\|_\infty \\ &\leq \kappa\, \sigma \sqrt{\log p} + 3\widetilde{\lambda}. \end{aligned}$$

Using the lower bound (3.11), and the expression of κ, we obtain

$$\|\widetilde{\beta}_T - \beta_T\|_\infty \leq 4\widetilde{\lambda}. \tag{3.3}$$

A sufficient condition to guarantee that the sign pattern is recovered is that this last upper bound be lower than the minimum absolute value of non-zero components of β, i.e.

$$4\widetilde{\lambda} \leq \min_{j \in T} |\beta_j|. \tag{3.4}$$

Using the upper bound on $\widetilde{\lambda}$ in (3.10), this is achieved in particular when

$$4\sigma\, \frac{\sqrt{n-s} + \sqrt{2\alpha\, \log(p)}}{\sqrt{s_0}}\, \frac{1-r}{\sqrt{1+r}} \leq \min_{j \in T} |\beta_j|,$$

which is implied by Assumption 2.5.

$\widetilde{\beta}$ and $\widetilde{\lambda}$ satisfy the optimality conditions

Using the lower bound (3.11) on $\widetilde{\lambda}$, the proof of the fact $\widetilde{\beta}$ and $\widetilde{\lambda}$ satisfy the optimality conditions is exactly the same as in [14, Section 3.5]. We repeat the argument for the sake of completeness. On one hand, by construction, we clearly have

$$X_T^t(y - X\widetilde{\beta}) = -\widetilde{\lambda}\, sign(\beta_T).$$

Since $\widetilde{\beta}$ and β have the same sign pattern, we actually have:

$$X_T^t(y - X\widetilde{\beta}) \quad = \quad -\widetilde{\lambda}\ sign(\widetilde{\beta}_T).$$

On the other hand,

$$\begin{aligned} \|X_{T^c}^t(y - X\widetilde{\beta})\|_\infty &= \|X_{T^c}^t \mathbf{P}_{V^\perp}(z) + \widetilde{\lambda}\ X_{T^c}^t X_T (X_T^t X_T)^{-1} sign(\beta_T)\|_\infty \\ &\leq \|X_{T^c}^t \mathbf{P}_{V^\perp}(z)\|_\infty + \widetilde{\lambda}\ \|X_{T^c}^t X_T (X_T^t X_T)^{-1} sign(\beta_T)\|_\infty \\ &\leq \kappa\ \sigma\sqrt{\log p} + \frac{1}{4}\widetilde{\lambda} \qquad (3.0) \\ &\leq \frac{3}{4}\ \widetilde{\lambda}\ < \ \widetilde{\lambda}. \end{aligned}$$

Hence, the two parts of the subgradient conditions (6.-35-6.-34) are satisfied by $\widetilde{\beta}$ and $\widetilde{\lambda}$, which means that

$$\widetilde{\beta} \quad = \quad \widehat{\beta}_{\widetilde{\lambda}}. \qquad (3.0)$$

In other words, $\widetilde{\beta}$ corresponds to the solution of problem (1.2) with the penalization $\lambda = \widetilde{\lambda}$. Moreover, $\widetilde{\lambda}$ bas been determined so that it verifies (3.6)

$$\widetilde{\lambda}^2 \quad = \quad C_{var}\ \frac{\|y - X_T\widetilde{\beta}_T\|_2^2}{n}\ \log p,$$

i.e., plugging (3.0),

$$\widetilde{\lambda}^2 \quad = \quad C_{var}\ \frac{\|y - X_T(\widehat{\beta}_{\widetilde{\lambda}})_T\|_2^2}{n}\ \log p.$$

Therefore, $\widetilde{\lambda}$ is a solution of Eq. (2.5). By virtue of uniqueness proved in Appendix 6, we deduce that

$$\begin{aligned} \widehat{\beta} &= \widetilde{\beta} \\ \widehat{\lambda} &= \widetilde{\lambda}. \end{aligned}$$

Conclusion of the proof

The two preceding sub-sections prove that $\widehat{\beta}$ has the same support and sign pattern as β. This occurs when (3.2) and (3.3) (both implied by the invertibility condition (v) in Sec. 3), Candès and Plan's conditions (i-iv) in Sec. 3 and the bound on $\|\mathbf{P}_{V_T^\perp} z\|_2$ in Sec. 3 are satisfied simultaneously. Therefore, this occurs with probability at least

$$1 - \frac{216 + 10 + 2}{p^\alpha},$$

as announced.

4 Proof of Theorem 2.7

As in the proof of Theorem 2.5, the quasi-isometry property (3.2) and (3.3), and Candès and Plan's conditions of Section 3 will be assumed. Notice also that the results of Section 6 are still valid with the assumption of Theorem 2.7.

The oracle estimator

As in the case of Section 3, the oracle for β is given by

$$\widetilde{\beta}_T = \left(X_T^t X_T\right)^{-1}\left(X_T^t y - \widetilde{\lambda}\ sign(\beta_T)\right). \tag{4.-4}$$

We now seek $\widetilde{\lambda}$ verifying

$$\frac{1}{2}\|y - X_T\widetilde{\beta}_T\|_2^2 = C\widetilde{\lambda}\ sign(\beta_T)^t\widetilde{\beta}_T. \tag{4.-3}$$

Replacing $\widetilde{\beta}$ by its value (3.4), we obtain

$$\begin{aligned} 1/2 \quad & \|y - X_T\left(X_T^t X_T\right)^{-1}\left(X_T^t y - \widetilde{\lambda}\ sign(\beta_T)\right)\|_2^2 \\ & = C\ \widetilde{\lambda}\ sign(\beta_T)^t\left(\left(X_T^t X_T\right)^{-1}\left(X_T^t y - \widetilde{\lambda}\ sign(\beta_T)\right)\right). \end{aligned}$$

Thus,

$$\begin{aligned} \frac{1}{2}\ \|\mathbf{P}_{V_T^\perp} y + \widetilde{\lambda} X_T\left(X_T^t X_T\right)^{-1} sign(\beta_T)\|_2^2 \ = \\ -C\widetilde{\lambda}^2\langle sign(\beta_T), \left(X_T^t X_T\right)^{-1} sign(\beta_T)\rangle + C\widetilde{\lambda}\ sign(\beta_T)^t\left(X_T^t X_T\right)^{-1} X_T^t y. \end{aligned}$$

Using the orthogonality relations, we then obtain

$$\begin{aligned} \frac{1}{2}\|\mathbf{P}_{V_T^\perp} y\|_2^2 + \frac{\widetilde{\lambda}^2}{2}\|X_T\left(X_T^t X_T\right)^{-1} sign(\beta_T)\|_2^2 = C\widetilde{\lambda}\ sign(\beta_T)^t\left(X_T^t X_T\right)^{-1} X_T^t y \\ -C\ \widetilde{\lambda}^2\langle sign(\beta_T), \left(X_T^t X_T\right)^{-1} sign(\beta_T)\rangle, \end{aligned}$$

which is equivalent to

$$\left(\frac{1}{2} + C\right)\widetilde{\lambda}^2\|\left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2^2 - C\widetilde{\lambda}\ sign(\beta_T)^t\left(X_T^t X_T\right)^{-1} X_T^t y + \frac{1}{2}\|\mathbf{P}_{V_T^\perp} z\|_2^2 = 0 \tag{4.-8}$$

The roots of the quadratic equation are

$$\widetilde{\lambda} = \frac{C\ sign(\beta_T)^t\left(X_T^t X_T\right)^{-1} X_T^t y \pm \sqrt{\Delta}}{(1+2C)\|\left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2^2}, \tag{4.-7}$$

where

$$\begin{aligned} \Delta = & \left(C\ sign(\beta_T)^t\left(X_T^t X_T\right)^{-1} X_T^t y\right)^2 \\ & -(1+2C)\|\left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2^2\|\mathbf{P}_{V_T^\perp} z\|_2^2. \end{aligned}$$

Study of the oracle $\tilde{\lambda}$

Following the same strategy as for Strategy (A), we now provide a confidence interval for $\tilde{\lambda}$.

Premilinaries

We have

$$\begin{aligned} sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y &= sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t (X_T\beta + z) \\ &= sign(\beta_T)^t \beta + sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t z \\ &= \|\beta\|_1 + \langle X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T), \mathbf{P}_{V_T} z + \mathbf{P}_{V_T^\perp} z\rangle. \end{aligned}$$

Hence,

$$sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y = \|\beta\|_1 + \langle X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T), \mathbf{P}_{V_T} z\rangle. \tag{4.-12}$$

Note that the Cauchy-Schwarz inequality yields

$$\left|\langle X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T), \mathbf{P}_{V_T} z\rangle\right| \leq \| \left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2 \|\mathbf{P}_{V_T} z\|_2. \tag{4.-12}$$

Bound on $\|\mathbf{P}_{V_T} z\|_2$

Using some well known properties of the χ^2 distribution recalled in Lemma 6.2 in the Appendix, we obtain

$$\mathbb{P}\left(\|\mathbf{P}_{V_T}(z)\|_2/\sigma \geq \sqrt{s} + \sqrt{2t}\right) \leq \exp(-t). \tag{4.-11}$$

Tune t such that $e^{-t} = 2p^{-\alpha}$, i.e.

$$t = \log(p^\alpha/2).$$

Hence,

$$\mathbb{P}\left(\|\mathbf{P}_{V_T}(z)\|_2/\sigma \geq \sqrt{s} + \sqrt{2\log(p^\alpha/2)}\right) \leq p^{-\alpha}. \tag{4.-11}$$

Positivity of Δ

We begin with the study of $sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y$ and $\| \left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2^2 \|\mathbf{P}_{V_T^\perp} z\|_2^2$, two key quantities in the analysis.

We first study $sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y$. By (3.2), we have

$$\| \left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2 \leq \sqrt{\frac{s}{1-r}}. \tag{4.-10}$$

Thus, using (4.-11), (4) and the lower bound (2.8) from Assumption 2.5 on the non-zero components of β, we can write

$$\begin{aligned} \left|\langle X_T \left(X_T^t X_T\right)^{-1} sign(\beta_T), \mathbf{P}_{V_T} z\rangle\right| &\leq \sigma \frac{\sqrt{s}}{\sqrt{1-r}} \left(\sqrt{s} + \sqrt{2\alpha \log p}\right) \\ &\leq \frac{1}{2} \|\beta\|_1. \end{aligned}$$

Therefore, from (4) we deduce that

$$\frac{1}{2}\|\beta\|_1 \leq \ sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y \leq \frac{3}{2}\|\beta\|_1. \tag{4.-11}$$

Second, we study $\| \left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2^2 \|\mathbf{P}_{V_T^\perp} z\|_2^2$. We have

$$\| \left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2 \|\mathbf{P}_{V_T^\perp} z\|_2 \quad \leq \quad \sigma\sqrt{\frac{s}{1-r}}\left(\sqrt{n-s}+\sqrt{2\alpha \log p}\right).$$

Thus

$$\begin{aligned} \Delta \quad &\geq \quad \frac{C^2}{4}\|\beta\|_1^2 - \sigma^2(1+2C)\frac{s}{1-r}\left(\sqrt{n-s}+\sqrt{2\alpha\log p}\right)^2 && (4.-12)\\ &\geq \quad \frac{C^2}{4}s^2 \min_{1\leq j\leq p}|\beta_j|^2 - \sigma^2(1+2C)\frac{s}{1-r}\left(\sqrt{n-s}+\sqrt{2\alpha\log p}\right)^2 && (4.-11) \end{aligned}$$

and Assumption 2.5 shows that $\Delta > 0$, which ensures that $\widetilde{\lambda}$ is well defined.

Bounds on $\tilde{\lambda}$

First, let us write

$$\sqrt{\Delta} = \left(C\ sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y\right) \times \sqrt{1 - \frac{(1+2C)\ \| \left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2^2 \|\mathbf{P}_{V_T^\perp} z\|_2^2}{\left(C\ sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y\right)^2}}.$$

On one hand, due to $\sqrt{1-\delta} \leq 1 - \frac{\delta}{2}$ on $(0,1)$, we obtain

$$\begin{aligned} \sqrt{\Delta} \quad \leq \quad &\left(C\ sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y\right) \\ &- \frac{(1+2C)\| \left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2^2 \|\mathbf{P}_{V_T^\perp} z\|_2^2}{2C\ sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y}. \end{aligned}$$

Combining this last equation with (4.-7), we obtain that

$$\widetilde{\lambda} \quad \geq \quad \frac{\|\mathbf{P}_{V_T^\perp} z\|_2^2}{2C\ sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y}. \tag{4.-14}$$

On the other hand, we also have $\sqrt{1-\delta} \geq 1 - \delta$ on $(0,1)$. Thus we can write

$$\begin{aligned} \sqrt{\Delta} \quad \geq \quad &\left(C\ sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y\right) \\ &- \frac{(1+2C)\| \left(X_T^t X_T\right)^{-\frac{1}{2}} sign(\beta_T)\|_2^2 \|\mathbf{P}_{V_T^\perp} z\|_2^2}{C\ sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y} \end{aligned}$$

and combining this last equation with (4.-7) and the previous upper bound, we thus obtain

$$\widetilde{\lambda} \quad \leq \quad \frac{\|\mathbf{P}_{V_T^\perp} z\|_2^2}{C\ sign(\beta_T)^t \left(X_T^t X_T\right)^{-1} X_T^t y}.$$

Using (4.-11), we finally get

$$\frac{\|\mathbf{P}_{V_T^\perp} z\|_2^2}{3\ C\ \|\beta\|_1} \ \leq\ \widetilde{\lambda}\ \leq\ 2\frac{\|\mathbf{P}_{V_T^\perp} z\|_2^2}{C\ \|\beta\|_1}. \tag{4.-16}$$

Combining this last equation with (3.9), we obtain:

$$\sigma^2\frac{(n-s)\left(\frac{\sqrt{\pi(n-s)}}{p^\alpha}\right)^{\frac{4}{n-s}}}{3\ C\ \|\beta\|_1} \ \leq\ \widetilde{\lambda}\ \leq\ 2\ \sigma^2\ \frac{\left(\sqrt{n-s}+\sqrt{2\alpha\log p}\right)^2}{C\|\beta\|_1}. \tag{4.-16}$$

Using Assumption 2.5 and (2.7), we thus obtain

$$\widetilde{\lambda}\ \ \geq\ \ \kappa\ \sigma\sqrt{\log p}. \tag{4.-15}$$

Last step of the proof

$\widetilde{\beta}$ and β have the same support and sign pattern

As in the case of Strategy (A) it is clear that $\tilde{\beta}$ and β have the same support. Let us now verify that they have the same sign pattern.

As in Section 3 and based on (4.-15), we obtain

$$\|\widetilde{\beta}_T - \beta_T\|_\infty\ \ \leq\ \ 4\widetilde{\lambda},$$

exactly as for Strategy (A). Using the upper bound on $\widetilde{\lambda}$ in the right hand side of (4.-16), we thus need

$$8\ \frac{\left(\sqrt{n-s}+\sqrt{2\alpha\log p}\right)^2}{C}\ \ \leq\ \ \min_{j\in T}|\beta_j|\ \frac{\|\beta\|_1}{\sigma^2}$$

to garantee that $\widetilde{\beta}_T$ and β_T have the same sign pattern. In view of this inequality, and since $\|\beta\|_1 \geq s\min_{j\in T}|\beta_j|$, an even stronger sufficient condition is

$$8\ \frac{\left(\sqrt{n-s}+\sqrt{2\alpha\log p}\right)^2}{C\ s}\ \ \leq\ \ \frac{\min_{j\in T}|\beta_j|^2}{\sigma^2}.$$

Noting that $\frac{2\sqrt{2}}{\sqrt{C}} \leq 2\frac{\sqrt{1+2C}}{C\sqrt{1-r}}$, we conclude that this condition is also implied by Assumption 2.5.

$\widetilde{\beta}$ and $\widetilde{\lambda}$ satisfy the optimality conditions

The proof is exactly the same as in Section 3 after replacing (3.11) by (4.-15).

Conclusion of the proof

The two preceding sub-sections prove that $\widehat{\beta}$ has same support and sign pattern as β. This occurs under the same conditions as those mentioned in the conclusion of the proof of Theorem 2.5, Sec. 3, plus the bound on $\|\mathbf{P}_{V_T} z\|_2$ in Sec. 4. Hence, this occurs with probability at least

$$1 - \frac{216+10+2+1}{p^\alpha},$$

as announced.

Epilogue: Nonempty range for $\|\beta\|_1$

We need to ensure that the range of admissible values for β is sufficiently large. The intuition says that this can be achieved by allowing sufficiently large values of n. In other words, we would like to know the additional constraints on the various parameters ensuring

$$s\, \mathsf{L}^{n,s,p}_{\alpha,r,C} \quad < \quad \mathsf{M}^{n,s,p}_{\alpha,r,\theta,C}.$$

Based on Eq. (2.6) and (2.7), it then suffices to know when the following inequalities are satisfied:

$$m_{\alpha,r,C}\ s\ \frac{\sqrt{n-s}+\sqrt{2\alpha\log p}}{\sqrt{s}} \quad \leq \quad \frac{n-s}{\sqrt{\log p}}\left(\frac{\sqrt{\pi(n-s)}}{p^{\alpha}}\right)^{\frac{4}{n-s}} \tag{4.-19}$$

where

$$m_{\alpha,r,C} \quad = \quad 6\kappa\frac{\sqrt{1+2C}}{\sqrt{1-r}}.$$

First, notice that under the condition

$$n-s \quad \geq \quad 8\alpha\ s\ \log p \geq 8\alpha\log p, \tag{4.-19}$$

we have $\log\left(\frac{\sqrt{\pi(n-s)}}{p^{\alpha}}\right)^{\frac{4}{n-s}} = \frac{4}{(n-s)}\left(\frac{1}{2}\left(\log(\pi)+\log(n-s)\right)-\alpha\log p\right) \geq -\frac{1}{2}$, and then

$$e^{-1/2} \quad \leq \quad \left(\frac{\sqrt{\pi(n-s)}}{p^{\alpha}}\right)^{\frac{4}{n-s}}.$$

Therefore, since we also have $\sqrt{2\alpha\log p} \leq \sqrt{n-s}$, (4.-19) is fulfilled if

$$2m_{\alpha,r,C}\sqrt{s}\ \sqrt{n-s} \quad \leq \quad e^{-1/2}\frac{n-s}{\sqrt{\log p}},$$

i.e.

$$n-s \quad \geq \quad 4e\ m^2_{\alpha,r,C}\ s\ \log p.$$

This explains the constraint (2.10) with the constant $c_\circ := 4e\ m^2_{\alpha,r,C} > 8\alpha$.

5 Algorithms and simulations results

In this section, we propose one iterative algorithm for Strategies (A) and (B) and we study their practical performance via Monte Carlo experiments.

We performed Monte Carlo experiments in the following setting. We took $p = 600$, $n = 75$ and $s = 9$ and we ran 500 experiments with $\sigma^2 = 1$ and the coefficients of β were randomly drawn independently as B times a Bernoulli ± 1 random variable plus an independent centered Gaussian perturbation with variance one.

Preliminaries

Our algorithms will be well defined under the assumption that for each positive value of the relaxation parameter, the value $\widehat{\beta}_\lambda$ of the regression vector is unique and the trajectory of $\widehat{\beta}_\lambda$ is continuous and piecewise affine. This property is well known under various assumptions on the design matrix X. It is a basic prerequisite for the theory behind Least Angle Regression and Homotopy methods. We refer the reader to [29] and [18] for information on these problems. See also [17] for a recent account on the study of $\widehat{\beta}_\lambda$ as a function of λ under generic conditions on the design matrix.

The subgradient conditions for the LASSO imply that

$$X^t_{\widehat{T}_\lambda}(y - X_{\widehat{T}_\lambda}\widehat{\beta}_{\widehat{T}_\lambda}) \quad = \quad \lambda\ sign(\widehat{\beta}_{\widehat{T}_\lambda}). \tag{5.-21}$$

where $X_{\widehat{T}_\lambda}$ is non-singular, and we obtain the well known fact that, for any $\lambda > 0$ such that $\widehat{\beta}_\lambda \neq 0$,

$$\widehat{\beta}_{\widehat{T}_\lambda} \quad = \quad (X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-1}\left(X^t_{\widehat{T}_\lambda} y - \lambda\ sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)\right). \tag{5.-20}$$

The following result is straightforward but useful.

Lemma 5.1 *(Nontriviality of the estimator) Let Σ be the set*

$$\Sigma = \left\{(S,\delta);\ S \subset \{1,\dots,p\},\ \delta \in \{-1,1\}^{|S|},\ |S| \leq n,\ \sigma_{\min}(X_S) > 0\right\}. \tag{5.-20}$$

The inequality

$$\inf_{(S,\delta)\in\Sigma} \left\|(X_S^t X_S)^{-1}(X_S^t y - \lambda\delta)\right\|_1 \quad > \quad 0 \tag{5.-19}$$

holds with probability one.

extbfProof. This is an immediate consequence of the Gaussian distribution of z. □

The standard LASSO with known variance

Simulations results: high SNR

With the choice $B = 40$, in all of the 500 experiments, we found that the support was exactly recovered.

Simulations results: low SNR

Figure H.1 below shows the histogram of the number of properly recovered components (left column) and the number of false components (right column) for the LASSO estimator with known variance and $\lambda = 2\sigma\sqrt{2\log p}$.

Strategy (A)

Implementation

As was discussed in Section 2, finding the estimator $(\widehat{\beta}, \widehat{\lambda})$ in Strategy A is equivalent to solving the equation

$$\Gamma_A(\lambda) \quad = \quad C_{var}.$$

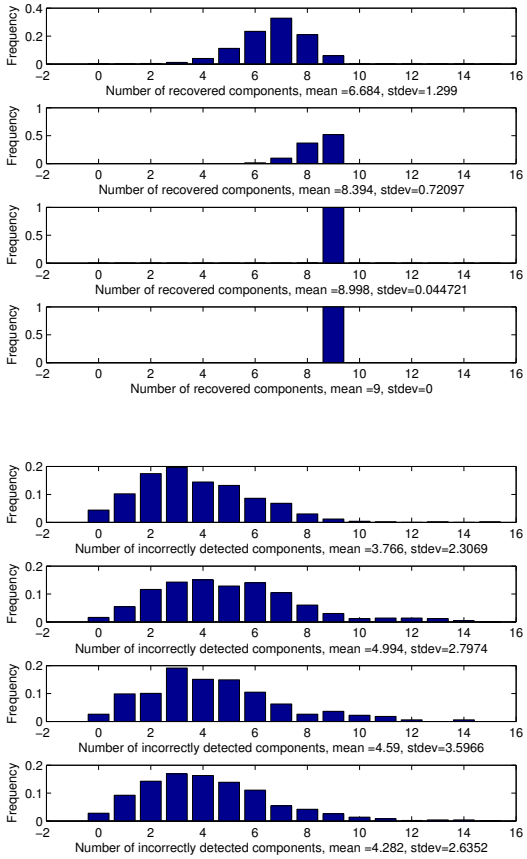

Figure H.1: Histogram of the number of properly recovered components (left column) and the number of false components (right column) for the LASSO estimator with known variance $\sigma^2 = 1$ and $\lambda = 2\sigma\sqrt{2\log p}$ for coeff. mean level $B = 1, 2, 5, 10$ (from top to bottom)

Since the function Γ_A is increasing (see Appendix 6.4), there is a number of Newton-type methods which can be used to solve this equation very efficiently and globally, i.e. without any condition on the initial iterate $\lambda^{(0)}$; see e.g. [32]. Instead of such refined methods, one may also use a simple dichotomic search. This is the option we chose for our simulations experiments, since the problem is only one dimensional and the complexity of the method is satisfactory in such simple cases. Notice that λ has the well-known upper bound $\|X^t y\|_\infty$ (a value beyond which $\widehat{\beta}_\lambda = 0$ [30]) and we used this value in the dichotomic search by specifying the initial interval to be $[0, \|X^t y\|_\infty]$.

Simulations results: high SNR

As for the case of the standard LASSO with known variance of Section 5 we found that, for $B = 40$, the support was exactly recovered in all of the 500 experiments.

Simulations results: low SNR

We performed Monte Carlo experiments in the same setting as for the LASSO in Section 5.

In real situations where the level of magnitude of the regression coefficients may not be much higher than the noise level, one observes that false positives often occur for the LASSO estimator with known variance. As seen from these results, the LASSO estimator where the variance is estimated using the penalty $\widehat{\lambda} = 2\widehat{\sigma}\sqrt{2\log p}$ performs at least as well as the standard LASSO estimator to which the true variance is available. The estimator $\widehat{\sigma}$ of the standard deviation is a slightly biased as shown in Figure H.3.

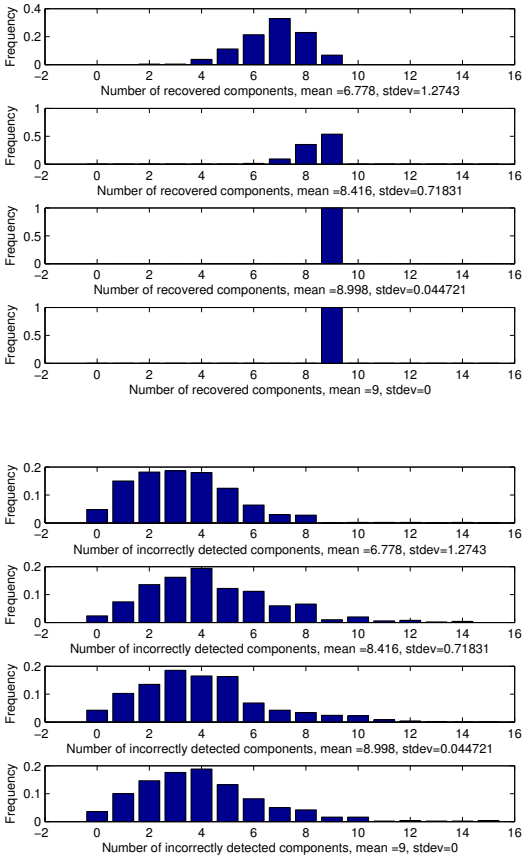

Figure H.2: Histogram of the number of properly recovered components (left column) and the number of false components (right column) for the LASSO estimator with unknown variance ($\sigma^2 = 1$) using Strategy (A) and $\widehat{\lambda} = 2\widehat{\sigma}\sqrt{2\log p}$ for coeff. mean level $B = 1, 2, 5, 10$ (from top to bottom).

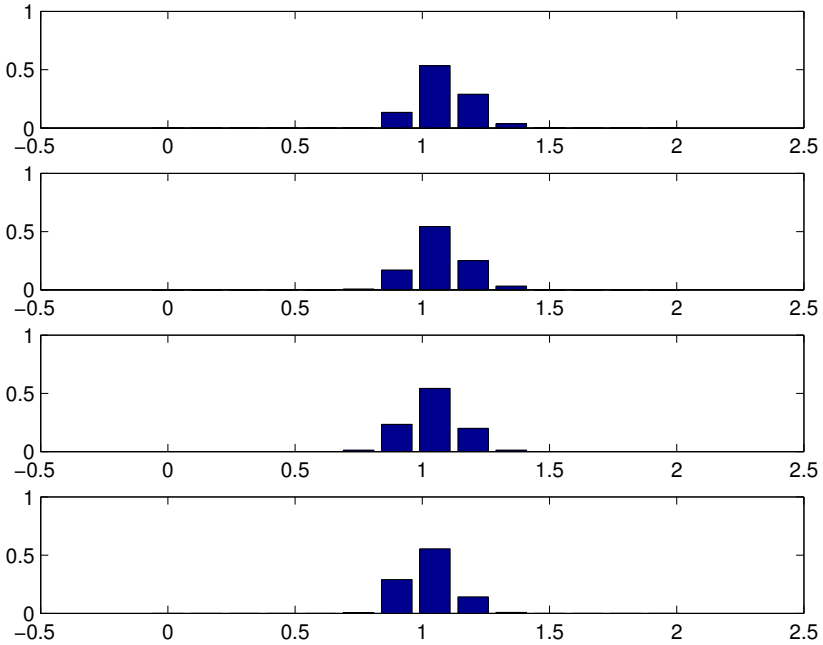

Figure H.3: Histogram of $\widehat{\sigma}$ for the LASSO estimator with unknown variance using Strategy (A) and $\widehat{\lambda} = 2\widehat{\sigma}\sqrt{2\log p}$ for coeff. mean level $B = 1, 2, 5, 10$ (from top to bottom).

Strategy (B)

Implementation

In order to compute the LASSO estimators $(\widehat{\beta}, \widehat{\lambda})$ satisfying the penalty vs. fidelity tradeoff constraint, we need to find $\widehat{\lambda}$ such that $\Gamma_B(\widehat{\lambda}) = C$. Since Γ is strictly decreasing by Lemma 6.5, and the problem is one dimensional, this task is not difficult to perform. As for Strategy A, we chose a standard dichotomic search for this problem. As for Strategy (A), we used the well-known upper bound $\|X^t y\|_\infty$ on λ (a value beyond which $\widehat{\beta}_\lambda = 0$) in the dichotomic search.

Simulations results: high SNR

As for the case of the standard LASSO with known variance of Section 5 we found that, for $B = 40$, the support was exactly recovered in all of the 100 experiments.

Simulations results: low SNR

We performed Monte Carlo experiments in the same setting as for the LASSO in Section 5.

Figure H.4 below shows the histogram of the number of properly recovered components (left column) and the number of false components (right column) for the LASSO estimator with unknown variance and the penalty vs. fidelity tradeoff constraint for the values $C =$

$.01, .1, .5, 5$. The instances where Newton's iterations did not converge were simply discarded although implementing a line search or a trust region strategy could easily have produced a correct result at the price of increasing the computational time for the Monte Carlo simulations study.

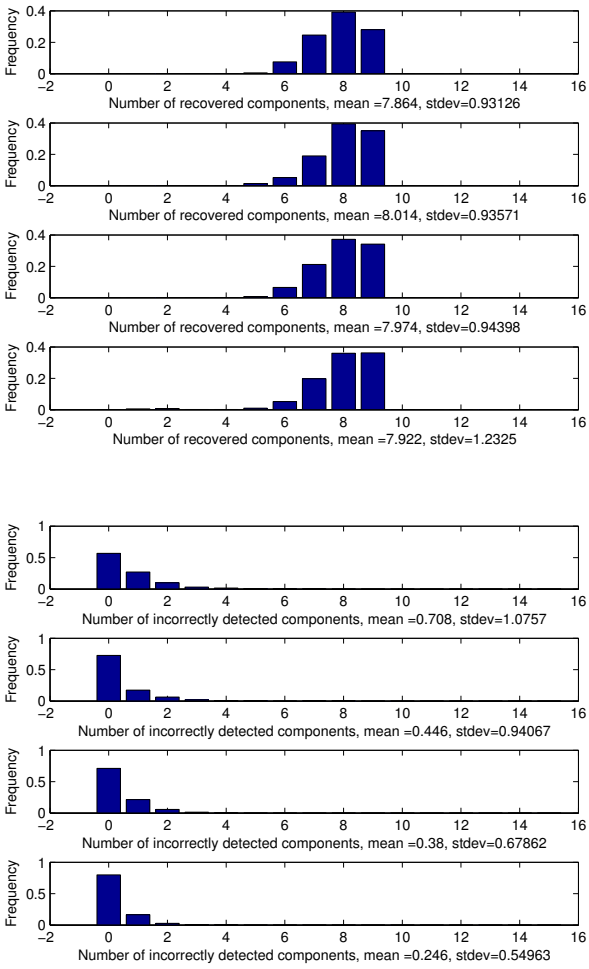

Figure H.4: Histogram of the number of properly recovered components (left column) and the number of false components (right column) for the LASSO estimator with unknown variance ($\sigma^2 = 1$) using Strategy (B) for $C = 0.01, .01, 0.5, 5$ (from top to bottom) with level $B = 2$.

The number of well recovered components of β is very close to the true value 9 for all values of C. On the other hand, the number of false positives is very close to zero for all values of C. Our estimator with penalty vs. fidelity tradeoff constraint is seen to have quite better performances than the standard LASSO and LASSO with estimated variance of the previous section with respect to the number of false positives; compare Figure H.4 with the second row of Figure H.1 or Figure H.2. This was the main objective for proposing this

strategy and the presented simulations show encouraging evidence of its robust behavior in the low SNR case. The low dependency on C is a property which might be well appreciated in practice when neither the signal nor the noise levels are precisely known ahead of time.

Comments

The simulations results confirmed the theoretical findings that, in the high SNR case, Strategy (A) and Strategy (B) perform as well, without knowing the variance ahead of time, as the standard LASSO which uses the true value of the variance. Although the results are presented for a particular set of parameters, this behavior was observed more generally for a large number of numerical experiments with different parameter configurations, for which the standard LASSO exactly recovers the true support and sign pattern. In the low SNR setting, the standard LASSO and Strategy (A) perform poorly in the sense that many false components are selected. The Monte Carlo experiments show that Strategy (B) is more robust in the low SNR setting, in the sense that the estimated support contains much less false components. Surprisingly, this phenomenon was observed over a wide range of values for the constant C. In other words, the dependence of Strategy (B)'s performance on C appeared as rather unessential for the recovery problem in the low SNR setting. As a preliminary practical conclusion, Strategy (A) appeared to be more suitable for the high SNR setting and Strategy (B) more suitable for the low SNR setting. In practice, the choice of C in Strategy B could be based on standard model selection procedures (AIC, BIC, Foster and George, etc) for comparing the obtained supports over a large range of possible values. The limited number of possible supports occurring in practice as C varies makes this comparison numerically tractable.

Finally, there remains the question of choosing between Strategy A and Strategy B on a given practical problem. One reasonable way to proceed might simply be as follows: compare the supports obtained via both methods, using a standard model selection procedure such as BIC, AIC, Foster and George's criterion, etc.

6 Appendices

Proof of Proposition 3.3

First, let us recall a technical result we obtained in [13]:

Lemma 6.1 *The following bound holds:*

$$\mathbb{P}\left(\|RH\|_{1\to 2} \geq v\right) \quad \leq \quad p\left(e\frac{s}{p}\frac{\|X\|^2}{v^2}\right)^{v^2/\mu(X)^2}, \tag{6.-19}$$

provided that $e\frac{s}{p}\frac{\|X\|^2}{v^2} \leq 1$.

Let us introduce the events:

$$\begin{aligned} E &= \{\|X_T^t X_T - I\| \leq r\} \\ B &= \left\{\|RH\|_{1\to 2} \leq \frac{c}{\sqrt{\log p}}\right\}. \end{aligned}$$

The proofs that Conditions (i) and (ii) hold with high probability are trivial modifications of the ones given in [14] up to the constants. The proofs that Conditions (iii) and (iv)

hold with high probability can be performed using the following by-product inequality from our Lemma 6.1:

$$\mathbb{P}\left(B^c\right) \quad \leq \quad p \exp\left(\frac{c^2}{C_\mu^2} \log\left(e \frac{C_{spar}}{c^2}\right) \log p\right), \tag{6.-20}$$

instead of using [14, Lemma 3.5] and [14, Lemma 3.6]. Here, we take

$$c^2 \quad \geq \quad \max(e^2 C_{spar} \, ; (1+\alpha) C_\mu), \tag{6.-19}$$

so that

$$\mathbb{P}\left(B^c\right) \quad \leq \quad \frac{1}{p^\alpha}. \tag{6.-18}$$

All the proofs are moreover based on the simple inequality

$$\begin{array}{rcl} \mathbb{P}(\mathcal{A}) & = & \mathbb{P}(\mathcal{A} \cap E \cap B) + \mathbb{P}(\mathcal{A} \cap (E^c \cup B^c)) \\ & \leq & E\left[\mathbb{P}\left(\mathcal{A} \mid R\right) \mathbb{I}_{E \cap B}\right] + \mathbb{P}(E^c) + \mathbb{P}(B^c), \end{array}$$

and the bound, for a given vector W:

$$\mathbb{P}\left(|\langle W, X\rangle| > t\right) \quad \leq \quad 2e^{-t^2/(2\|W\|_2^2)}. \tag{6.-19}$$

This last bound holds true for sub-Gaussian random vectors with independent components having Bernoulli or standard Gaussian distribution, for instance.

Condition (i)

Here, let W_i be the ith row of $(X_T^t X_T)^{-1} X_T^t$. Since $\langle W_i, z\rangle \sim \mathcal{N}(0, \|W_i\|_2^2)$, we have from (4.-76) and the union bound:

$$\mathbb{P}\left(\max_{i \in T} |\langle W_i, z\rangle| > t\right) \quad \leq \quad 2s \; e^{-t^2/(2 \max_i \|W_i\|_2^2)}.$$

Note that on E:

$$\max_{i \in T} \|W_i\|_2 \; \leq \; \|(X_T^t X_T)^{-1}\| \; \|X_T^t\| \; \leq \frac{\sqrt{1+r}}{1-r}. \tag{6.-20}$$

One then obtains

$$\mathbb{P}\left(\|(X_T^t X_T)^{-1} X_T^t z\|_\infty \leq \sigma \; \kappa \sqrt{\log p}\right) \quad \geq \quad 1 - \frac{2}{p^\alpha},$$

whenever

$$\kappa \quad \geq \quad \frac{\sqrt{2(1+\alpha)(1+r)}}{1-r}. \tag{6.-20}$$

Condition (ii)

Let us show that the estimate (ii) holds with high probability. This is an actual consequence of our Lemma 6.1.

First, as in [14] p.2171 and Lemma 3.3 p.2166, we write the inequality

$$\|(X_T^t X_T)^{-1} sign(\beta_T)\|_\infty \quad \leq \quad 1 + \max_{i \in T} |\langle W_i, sign(\beta_T)\rangle|,$$

where W_i is the ith row or column of $(X_T^t X_T)^{-1} - I$. Set

$$\mathcal{A} \quad = \quad \left\{ \max_{i \in J} |\langle W_i, sign(\beta_T) \rangle| \geq 2 \right\}.$$

Hoeffding's inequality yields:

$$\mathbb{P}\left(\mathcal{A}|R\right) \ \leq \ 2|J| \exp\left(- \frac{2^2}{2 \max_{i \in J} \|W_i\|_2^2} \right). \tag{6.-22}$$

As in [14] p.2171 and p.2172, we write $\|W_i\|_2 \leq \frac{\|RHRe_i\|_2}{1-r}$. Thus on E:

$$\|W_i\|_2 \leq \frac{\|RHR\|_{1\to 2}}{1-r} \leq \frac{\|RH\|_{1\to 2}}{1-r}.$$

Recall that $\mathbb{P}(B^c) \leq \frac{1}{p^\alpha}$ since c satisfies (6.-19). Moreover

$$E\left[\mathbb{P}\left(\mathcal{A} \mid R\right) \mathbb{I}_{E \cap B}\right] \quad \leq \quad \frac{1}{p^\alpha}$$

holds true if

$$c^2 \quad \leq \quad \frac{2(1-r)}{1+\alpha}.$$

We can easily check that this last condition is compatible with (6.-19) and

$$\begin{aligned} C_\mu \quad &= \quad \frac{r}{1+\alpha} \\ C_{spar} \quad &= \quad \frac{r^2}{(1+\alpha)e^2}, \end{aligned}$$

whenever $r \in (0, 1/2)$. Therefore, when $r \in (0, 1/2)$, the event

$$\|(X_T^t X_T)^{-1} sign(\beta_T)\|_\infty \leq 1 + 2 = 3$$

holds with probability at least $1 - \frac{3}{p^\alpha}$.

Condition (iii)

Here, $W_i = (X_T^t X_T)^{-1} X_T^t X_i$. Notice that on $E \cap B$:

$$\max_{i \in T^c} \|W_i\|_2 \ \leq \frac{c}{(1-r)\sqrt{\log p}}. \tag{6.-27}$$

Using (6.-18) again and the same previous arguments, we obtain

$$\mathbb{P}\left(\|X_{T^c}^t X_T (X_T^t X_T)^{-1} sign(\beta_T)\|_\infty \leq \frac{1}{4} \right) \quad \geq \quad 1 - \frac{3}{p^\alpha}.$$

Condition (iv)

If one now sets W_i as the ith row of $I - X_T(X_T^t X_T)^{-1} X_T^t$ and note that on E for any $i \in T$:

$$\|W_i\|_2 \leq \|X_i\|_2 = 1, \tag{6.-28}$$

then:

$$\mathbb{P}\left(\|X_{T^c}^t \left(I - X_T (X_T^t X_T)^{-1} X_T^t\right) z\|_\infty \leq \sigma \ \kappa \ \sqrt{\log p} \right) \quad \geq \quad 1 - \frac{2}{p^\alpha},$$

whenever

$$\kappa \quad \geq \quad \sqrt{2(1+\alpha)}. \tag{6.-28}$$

Choosing κ

The parameter κ has to satisfy (6.-20) and (6.-28). Since $r \in (0, \frac{1}{2}]$, one has $\frac{\sqrt{2(1+r)}}{1-r} \leq 2\sqrt{3} \approx 3.4$. Thus we simply chose

$$\kappa \quad = \quad 4\sqrt{1+\alpha},$$

which is Eq. (2.4).

Some properties of the χ^2 distribution

We recall the following useful bounds for the $\chi^2(\nu)$ distribution of degree of freedom ν.

Lemma 6.2 *The following bounds hold:*

$$\begin{aligned} \mathbb{P}\left(\chi(\nu) \geq \sqrt{\nu} + \sqrt{2t}\right) &\leq \exp(-t) \\ \mathbb{P}\left(\chi(\nu) \leq \sqrt{u\nu}\right) &\leq \frac{2}{\sqrt{\pi\nu}} \left(u\ e/2\right)^{\frac{\nu}{4}}. \end{aligned}$$

extbfProof. For the first statement, see e.g. [27]. For the second statement, recall that

$$\begin{aligned} \mathbb{P}\left(\chi^2(\nu) \leq u\nu\right) &= \int_0^{u\frac{\nu}{2}} \frac{t^{\frac{\nu}{2}-1}e^{-t}}{\Gamma(\frac{\nu}{2})} dt \\ &= \int_0^{u\frac{\nu}{2}} \frac{t^{\frac{\nu}{2}-1-\alpha}t^{\alpha}e^{-t}}{\Gamma(\frac{\nu}{2})} dt. \end{aligned}$$

Since $\max_{t\in\mathbb{R}^+} t^{\alpha}e^{-t} = (\alpha/e)^{\alpha}$ and is attained at $t = \alpha$, we obtain that

$$\mathbb{P}\left(\chi^2(\nu) \leq u\nu\right) \leq \frac{(\alpha/e)^{\alpha}}{\Gamma(\frac{\nu}{2})} \int_0^{u\frac{\nu}{2}} t^{\frac{\nu}{2}-1-\alpha} dt = \frac{(\alpha/e)^{\alpha}}{(\frac{\nu}{2}-\alpha)\Gamma(\frac{\nu}{2})} \left(u\frac{\nu}{2}\right)^{\frac{\nu}{2}-\alpha}.$$

Take for instance $\alpha = \frac{\nu}{4}$ and obtain

$$\mathbb{P}\left(\chi^2(\nu) \leq u\nu\right) = \frac{(\nu/4e)^{\frac{\nu}{4}}}{\frac{\nu}{4}\Gamma(\frac{\nu}{2})} \left(u\frac{\nu}{2}\right)^{\frac{\nu}{4}}. \tag{6.-33}$$

On the other hand, we have

$$\Gamma(z) \geq \sqrt{2\pi}\, \frac{z^{z-\frac{1}{2}}}{e^z}$$

and then,

$$\frac{(\nu/4e)^{\frac{\nu}{4}}}{\frac{\nu}{4}\Gamma(\frac{\nu}{2})} \leq \sqrt{\frac{2}{\pi}} \frac{(e/2)^{\frac{\nu}{4}} \left(\frac{\nu}{2}\right)^{-\frac{\nu}{4}}}{\sqrt{\frac{\nu}{2}}}.$$

Hence,

$$\mathbb{P}\left(\chi^2(\nu) \leq u\nu\right) \leq \sqrt{\frac{2}{\pi}} \frac{(u\ e/2)^{\frac{\nu}{4}}}{\sqrt{\frac{\nu}{2}}} = \frac{2}{\sqrt{\pi\nu}} \left(u\ e/2\right)^{\frac{\nu}{4}},$$

as desired. □

Properties of the standard LASSO

Reminders on the LASSO subgradient conditions

In [20] Section III, it is proven that a necessary and sufficient optimality condition in (1.2) is the two following conditions:

$$X_T^t(y - X\widehat{\beta}_\lambda) = \lambda\, sign(\beta_T) \quad (6.\text{-}35)$$

$$\|X_{T^c}^t(y - X\widehat{\beta}_\lambda)\|_\infty \leq \lambda. \quad (6.\text{-}34)$$

Moreover, if $\|X_{T^c}^t(y - X\widehat{\beta}_\lambda)\|_\infty < \lambda$, then problem (1.2) admits a unique solution.

Let us also recall (see [17] and [14]) that the support $\widehat{T}_\lambda \subset \{1, \dots, p\}$ of $\widehat{\beta}_\lambda$ verifies

$$|\widehat{T}_\lambda| \leq n. \quad (6.\text{-}33)$$

General properties of $\lambda \mapsto \widehat{\beta}_\lambda$

Recall that $\widehat{\beta}_\lambda$ is the standard LASSO estimator of β parametrized by λ,

The following notations will be useful. Define $\mathcal{L}$ as the cost function:

$$\mathcal{L} : \begin{cases} (0, +\infty) \times \mathbb{R}^p & \longrightarrow & \mathbb{R}_+ \\ (\lambda, b) & \longmapsto & \frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1, \end{cases} \quad (6.\text{-}33)$$

and for all $\lambda > 0$,

$$\theta(\lambda) = \inf_{b \in \mathbb{R}^p} \mathcal{L}(\lambda, b).$$

Lemma 6.3 *Let the Generic Condition hold. Then, the function θ is concave and non-decreasing.*

extbfProof. Since θ is the infimum of a set of affine functions of the variable λ, it is concave. Moreover, we have

$$\theta(\lambda) = \mathcal{L}(\lambda, \widehat{\beta}_\lambda),$$

where, by Proposition 2.2, $\widehat{\beta}$ is the unique solution of (1.0). Using the filling property [22, Chapter XII], we obtain that $\partial\theta(\lambda)$ is the singleton $\{\|\widehat{\beta}_\lambda\|_1\}$. Thus, θ is differentiable and its derivative at λ is given by

$$\theta'(\lambda) = \|\widehat{\beta}_\lambda\|_1.$$

Moreover, this last expression shows that θ is nondecreasing. □

Proof of Lemma 2.3

(i) $\|\widehat{\beta}_\lambda\|_1$ is non-increasing – The fact that $\lambda \longmapsto \|\widehat{\beta}_\lambda\|_1$ is non-increasing is an immediate consequence of the concavity of θ.

(ii) Boundedness – Notice that using (5.-20), we obtain that

$$\|\widehat{\beta}_\lambda\|_1 \leq \max_{(S,\delta)\in\Sigma} \left\|(X_S^t X_S)^{-1}(X_S^t y - \lambda\delta)\right\|_1,$$

where we recall that, in Lemma 5.1, Σ reads as

$$\Sigma \quad = \quad \left\{(S,\delta);\ S\subset\{1,\dots,p\},\ \delta\in\{-1,1\}^{|S|},\ |S|\leq n,\ \sigma_{\min}(X_S)>0\right\}.$$

Thus, $\lambda \longmapsto \widehat{\beta}_\lambda$ is bounded on any interval of the form $(0,M]$, with $M\in(0,+\infty)$. Moreover, since its ℓ_1-norm is non-increasing, it is bounded on $(0,\infty)$.

(iii) Continuity – Assume for contradiction that $\lambda \longmapsto \widehat{\beta}_\lambda$ is not continuous at some $\lambda^\circ>0$. Using boundedness, we can construct two sequences converging towards $\widehat{\beta}^+_{\lambda^\circ}$ and $\widehat{\beta}^-_{\lambda^\circ}$ respectively with $\widehat{\beta}^+_{\lambda^\circ}\neq\widehat{\beta}^-_{\lambda^\circ}$. Since $\mathcal{L}(\lambda^\circ,\cdot)$ is continuous, both limits are optimal solutions of the problem

$$\underset{b\in\mathbb{R}^p}{argmin}\ \mathcal{L}(\lambda^\circ,b), \tag{6.-37}$$

hence contradicting the uniqueness.

$\widehat{T}_\lambda$ **h**as cardinality n **f**or λ **s**ufficiently small.

Let $(\lambda_k)_{k\in\mathbb{N}}$ be any positive sequence converging to 0. Let β^* be any cluster point of the sequence $(\widehat{\beta}_{\lambda_k})_{k\in\mathbb{N}}$ (this sequence is easily seen to be bounded under various standard assumptions; see e.g. [14, Lemma 3.5] for a proof). Fix $\varepsilon>0$ and $b\in\mathbb{R}^p$. For all $k\in\mathbb{N}$, we have

$$\mathcal{L}(\lambda_k,\widehat{\beta}_{\lambda_k}) \quad\leq\quad \mathcal{L}(\lambda_k,b), \tag{6.-36}$$

where $\mathcal{L}$ is defined by (6). Since $\mathcal{L}(\lambda_k,\cdot)$ is continuous, we can also write for k sufficiently large:

$$\mathcal{L}(\lambda_k,\beta^*) \quad\leq\quad \mathcal{L}(\lambda_k,\widehat{\beta}_{\lambda_k})+\varepsilon.$$

Hence, $\mathcal{L}(\lambda_k,\beta^*)\leq\mathcal{L}(\lambda_k,b)+\varepsilon$. Letting $\lambda_k\to 0$, we obtain

$$\frac{1}{2}\|y-X\beta^*\|_2^2 \quad\leq\quad \frac{1}{2}\|y-Xb\|_2^2+\varepsilon,$$

and thus,

$$\frac{1}{2}\|y-X\beta^*\|_2^2 \quad\leq\quad \inf_{b\in\mathbb{R}^p}\frac{1}{2}\|y-Xb\|_2^2. \tag{6.-37}$$

Since $range(X)=\mathbb{R}^n$, (6.-37) implies $\|y-X\beta^*\|_2^2=0$, and then

$$\lim_{\lambda\downarrow 0}\|y-X\widehat{\beta}_\lambda\|_2^2 \quad=\quad 0. \tag{6.-36}$$

Notice further that $\{b\in\mathbb{R}^p,\ |supp(b)|<n\}$ is a finite union of subspaces of $\mathbb{R}^p$, each with dimension $n-1$. Thus,

$$m:=\inf_{\{b\in\mathbb{R}^p;\ |supp(b)|<n\}}\frac{1}{2}\|y-Xb\|_2^2 \quad>\quad 0,$$

with probability one. Therefore for λ sufficiently small, (6.-36) implies $\|y-X\widehat{\beta}_\lambda\|_2^2<m$, from which we deduce that $|\widehat{T}_\lambda|=n$ since one has $|\widehat{T}_\lambda|\leq n$ (cf Reminder 6).

Partitioning $(0,+\infty)$ into good intervals

The continuity of $\lambda\mapsto\widehat{\beta}_\lambda$ implies that the interval $(0,+\infty)$ can be partitioned into subintervals of the type $I_k=(\lambda_k,\lambda_{k+1}]$, with

(i) $\lambda_0 = 0$ and $\lambda_k \in (0, +\infty]$ for $k > 0$,

(ii) the support and sign pattern of $\widehat{\beta}_\lambda$ are constant on each open interval $\mathring{I}_k := (\lambda_k, \lambda_{k+1})$.

Notice further that due to 6, $\widehat{T}_\lambda \neq \emptyset$ on at least I_0. Let $\mathcal{K}$ be the nonempty set

$$\mathcal{K} \quad = \quad \left\{ k \in \mathbb{N},\ \forall \lambda \in \mathring{I}_k,\ \widehat{\beta}_\lambda \neq 0 \right\}.$$

On any interval I_k, $k \in \mathcal{K}$, uniqueness of $\widehat{\beta}$ implies that the expression (5.-20) for $\widehat{\beta}_{\widehat{T}_\lambda}$ holds. Multiplying (5.-20) on the left by $sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)^t$, we obtain

$$\|\widehat{\beta}_\lambda\|_1 = sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)^t (X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-1} X_T^t y - \lambda sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)^t (X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right).$$

Thus

$$\frac{d\|\widehat{\beta}_\lambda\|_1}{d\lambda}(\lambda) \quad = \quad -sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)^t (X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right),$$

on $(0, +\infty)$. Thus, the definition of Σ, we obtain that

$$\frac{d\|\widehat{\beta}_\lambda\|_1}{d\lambda}(\lambda) \quad \leq \quad -\inf_{(S,\delta)\in\Sigma} \delta^t (X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-2} \delta \ < \ 0 \tag{6.-39}$$

on each $\mathring{I}_k$, $k \in \mathcal{K}$ and

$$\frac{d\|\widehat{\beta}_\lambda\|_1}{d\lambda}(\lambda) \quad = \quad 0$$

on each $\mathring{I}_k$, $k \notin \mathcal{K}$, i.e. on each $\mathring{I}_k$ such that $\|\widehat{\beta}_{\widehat{T}_\lambda}\|_1 = 0$ for all λ in I_k, if any such I_k exists. Since $\lambda \longmapsto \|\widehat{\beta}_\lambda\|_1$ is continuous on $(0, \infty)$, (6.-39) implies that:

(i) there exists $\tau \in (0, +\infty)$, such that $\widehat{\beta}_\tau = 0$ (as an easy consequence of the Fundamental Theorem of Calculus and a contradiction).

(ii) $\widehat{\beta}_\lambda = 0$ for all $\lambda \geq \tau$.

Hence $\cup_{k\in\mathcal{K}} I_k$ is a connected bounded interval.

The map $\lambda \mapsto \|y - X\widehat{\beta}_\lambda\|_2$ is increasing on $(0, \tau]$

Using (5.-20), we obtain

$$y - X\widehat{\beta}_\lambda \quad = \quad P_{V^\perp_{\widehat{T}_\lambda}}(y) - \lambda X_{\widehat{T}_\lambda} (X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right),$$

which implies that

$$\begin{aligned} \|y - X\widehat{\beta}_\lambda\|_2^2 \quad = \quad & \left\| P_{V^\perp_{\widehat{T}_\lambda}}(y) \right\|_2^2 - 2\lambda \langle P_{V^\perp_{\widehat{T}_\lambda}}(y), X_{\widehat{T}_\lambda} (X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right) \rangle \\ & + \lambda^2\ sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)^t (X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right) \end{aligned}$$

and thus, by the definition of $P_{V^\perp_{\widehat{T}_\lambda}}(y)$,

$$\|y - X\widehat{\beta}_\lambda\|_2^2 \quad = \quad \left\| P_{V^\perp_{\widehat{T}_\lambda}}(y) \right\|_2^2 + \lambda^2\ sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)^t (X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right). \tag{6.-42}$$

From (6.-42), since $(X^t_{\widehat{T}_\lambda} X_{\widehat{T}_\lambda})^{-1}$ is definite, we obtain that $\lambda \mapsto \|y - X\widehat{\beta}_\lambda\|_2$ is increasing on each $\mathring{I}_k$, and thus on $(0, \tau]$ by using that $\lambda \mapsto \|y - X\widehat{\beta}_\lambda\|_2$ is continuous on $(0, \tau]$.

Study of Γ_A

Lemma 6.4 *Γ_A is increasing on $(0,\tau]$ and $\lim_{\lambda\to+\infty}\Gamma_A(\lambda)=+\infty$.*

extbfProof. Due to Step 3, and the definition of τ, the set of values $\lambda>0$ such that $\|y-X\widehat{\beta}_\lambda\|_2>0$ is nonempty. Let λ_{inf} denote its infimum value. Take $\lambda\in\mathring{I}_k$ for some k such that $\lambda\geq\lambda_{inf}$. In particular, $\lambda\neq 0$. Then,

$$\Gamma_A(\lambda) \;=\; \frac{s}{\frac{1}{\lambda^2}\left\|P_{V^\perp_{\widehat{T}_\lambda}}(y)\right\|_2^2+sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)^t(X^t_{\widehat{T}_\lambda}X_{\widehat{T}_\lambda})^{-1}sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)}, \tag{6.-41}$$

and we deduce that Γ_A is increasing on $\mathring{I}_k$. By continuity, we have that Γ_A is increasing on $(\lambda_{inf},\tau]$. Once $\lambda>\tau$, $\|y-X\widehat{\beta}_\lambda\|_2^2=\|y\|_2^2$ and $\Gamma_A(\lambda)=s\lambda^2/\|y\|_2^2$. Thus, $\lim_{\lambda\to+\infty}\Gamma_A(\lambda)=+\infty$ as desired. □

The fact that Γ_A is increasing proves that the equation $\Gamma_A(\lambda)=C_{var}$ admits at most one solution.

Study of Γ_B

Recall that

$$\Gamma_B(\lambda) \;=\; \frac{\lambda\|\widehat{\beta}_\lambda\|_1}{\|y-X\widehat{\beta}_\lambda\|_2^2}. \tag{6.-40}$$

We will use repeatedly that $\widehat{\beta}_\lambda$ is unique for all $\lambda>0$ and that the trajectory $\lambda\mapsto\widehat{\beta}_\lambda$ is continuous under the Generic Condition, see [17].

Lemma 6.5 *Under the Generic Position Assumption of [17], the function Γ_B defined by (6.-40) almost surely satisfies*

$$\lim_{\lambda\downarrow 0}\Gamma_B(\lambda) \;=\; +\infty. \tag{6.-39}$$

Moreover, almost surely, there exists $\tau>0$ such that Γ_B is decreasing on the interval $(0,\tau]$ with $\Gamma_B(\tau)=0$, while $\|y-X\widehat{\beta}_\lambda\|_2$ is increasing on $(0,\tau]$.

extbfProof. Let us first show that $\lim_{\lambda\downarrow 0}\Gamma_B(\lambda)=+\infty$.

Let $\lambda_0>0$ be sufficiently small so that for all $\lambda\leq\lambda_0$, $|\widehat{T}_\lambda|=n$. Such a λ_0 exists due to Step 1.a. Hence, since $X_{\widehat{T}_\lambda}$ is nonsingular:

$$\mathbf{P}_{V_{T_\lambda}} \;=\; I_n. \tag{6.-38}$$

Thus, using (5.-20), we obtain

$$y-X\widehat{\beta}_\lambda \;=\; -\lambda X_{\widehat{T}_\lambda}(X^t_{\widehat{T}_\lambda}X_{\widehat{T}_\lambda})^{-1}sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right), \tag{6.-37}$$

which implies that

$$\|y-X\widehat{\beta}_\lambda\|_2^2 \;=\; \lambda^2\|(X^t_{\widehat{T}_\lambda}X_{\widehat{T}_\lambda})^{-1}sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)\|_2^2.$$

Moreover, Lemma 5.1 combined with (5.-20) gives

$$\|\widehat{\beta}_\lambda\|_1 \quad > \inf_{(S,\delta)\in\Sigma} \left\|(X_S^t X_S)^{-1}(X_S^t y - \lambda\delta)\right\|_1 := m' > \quad 0.$$

Hence, for $\lambda \leq \lambda_0$,

$$\Gamma_B(\lambda) \quad \geq \quad \frac{\lambda m'}{\lambda^2 \|X_{\widehat{T}_\lambda}(X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)\|_2^2}.$$

Using the trivial fact that $\sup_{(S,\delta)\in\Sigma} \|X_S(X_S^t X_S)^{-1}\delta\|_2^2 < \infty$, the proof of Step 1 is complete.

Let us now show that Γ_B is decreasing on $(0,\tau)$ by studying the function

$$\Phi \ : \ \left\{ \begin{array}{ccc} (0,+\infty) & \longrightarrow & \mathbb{R}_+ \\ \lambda & \longmapsto & \lambda\|\widehat{\beta}_\lambda\|_1. \end{array} \right. \tag{6.-40}$$

We immediately deduce from Step 2 and the definition of the intervals I_k, $k \in \mathcal{K}$, that Φ is differentiable on each $\mathring{I}_k$, $k \in \mathcal{K}$. Using (5.-20), its derivative on $\mathring{I}_k$ reads

$$\begin{aligned} \frac{d\Phi}{d\lambda}(\lambda) &= \|\widehat{\beta}_{\widehat{T}_\lambda}\|_1 - \lambda \ sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)^t (X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right) \\ &= \|\widehat{\beta}_{\widehat{T}_\lambda}\|_1 - \lambda \ \|(X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1/2} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)\|_2^2. \end{aligned}$$

Now, since $X_{\widehat{T}_\lambda}$ is non singular,

$$\|y - X\widehat{\beta}_\lambda\|_2^2 = \lambda^2 \|(X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)\|_2^2 > \lambda^2 n \ \sigma_{\min}\left((X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1}\right)^2 > 0$$

for $\lambda > 0$. Therefore $\Gamma_B(\lambda) < +\infty$ on $(0,+\infty)$, Γ_B is continuous on I_k and differentiable on $\mathring{I}_k$. Moreover, using (6.-42), we have

$$\frac{d\Gamma_B}{d\lambda}(\lambda) = \frac{\frac{d\Phi}{d\lambda}(\lambda)\|y - X\widehat{\beta}_\lambda\|_2^2 - \Phi(\lambda)\frac{d\|y - X\widehat{\beta}_\lambda\|_2^2}{d\lambda}(\lambda)}{\|y - X\widehat{\beta}_\lambda\|_2^4} = \frac{\frac{d\Phi}{d\lambda}(\lambda) - 2\frac{\Phi(\lambda)}{\lambda}}{\|y - X\widehat{\beta}_\lambda\|_2^2}.$$

Hence, using (6.-40) and (6.-42),

$$\begin{aligned} \frac{d\Gamma_B}{d\lambda}(\lambda) &= \frac{-\|\widehat{\beta}_{\widehat{T}_\lambda}\|_1 - \lambda \ \|(X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1/2} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)\|_2^2}{\|y - X\widehat{\beta}_\lambda\|_2^2} \\ &\leq \frac{-\lambda \ \|(X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1/2} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)\|_2^2}{\lambda^2 \|(X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1} sign\left(\widehat{\beta}_{\widehat{T}_\lambda}\right)\|_2^2} \leq -\frac{1}{\lambda}\left(\frac{\sigma_{\min}\left((X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1/2}\right)}{\sigma_{\max}\left((X_{\widehat{T}_\lambda}^t X_{\widehat{T}_\lambda})^{-1}\right)}\right)^2, \end{aligned}$$

on each $\mathring{I}_k$. We can thus conclude, due to the non-singularity of $X_{\widehat{T}_\lambda}$, that Γ_B is decreasing on $(0,\tau)$, as announced.

□

Bibliography

[1] Bajwa, W.; Calderbank, R.; Mixon, D. Two are better than one: fundamental parameters of frame coherence. Appl. Comput. Harmon. Anal. 33 (2012), no. 1, 58–78. (p. 187).

[2] Baraud, Y., Giraud, C., Huet, S. Gaussian model selection with an unknown variance. Ann. Statist. 37 (2009), no. 2, 630–672. (p. 177).

[3] Bickel, P. J., Ritov, Y., Tsybakov, A. B., Simultaneous analysis of lasso and Dantzig selector. Ann. Statist. 37 (2009), no. 4, 1705–1732. (pp. 177 et 250).

[4] Bourgain, J., Tzafriri, L., Invertibility of "large" submatrices with applications to the geometry of Banach spaces and harmonic analysis. Israel J. Math. 57 (1987), no. 2, 137–224. (pp. 217 et 222).

[5] Belloni, A., Chernozhukov, V. and Wang, L., Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika 98 (2011), no. 4, 791–806. (p. 178).

[6] Städler, N., Bühlmann, P., and van de Geer, S. (2010), ℓ_1-penalization for mixture regression models, Test, 19, 209–285 (p. 178).

[7] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007a). Sparsity oracle inequalities for the Lasso. Electron. J. Stat., 1 :169–194. (p. 177).

[8] Candès, E. J. The restricted isometry property and its implications for compressed sensing. C. R. Math. Acad. Sci. Paris 346 (2008), no. 9-10, 589–592. (p. 250).

[9] Candès, E. J. Modern statistical estimation via oracle inequalities. Acta Numer. 15 (2006), 257–325.

[10] Candès, E. J. and Plan, Y. Near-ideal model selection by ℓ_1 minimization. Ann. Statist. 37 (2009), no. 5A, 285–2177. (pp. 177, 178, 180, 181, 182, 183, 184, 185, 187, 188, 189, 192, 193, 205, 206, 207, 217, 218, 219, 220, 228, 229, 230, 233, 243, 245, 250, 254, 255, 256, 286, 300 et 301).

[11] Candès, E. and Romberg, J., Sparsity and incoherence in compressive sampling. Inverse Problems 23 (2007), no. 3, 969–985. (p. 177).

[12] Candès, E. J. and Tao, T., The Dantzig Selector: statistical estimation when p is much larger than n. Ann. Stat. 35, no. 6 (2007), 2313–2351. (pp. 177 et 250).

[13] Chrétien, S. and Darses, S., Invertibility of random submatrices via tail decoupling and a Matrix Chernoff Inequality. Statist. Probab. Lett. 82 (2012), no. 7, 1479-1487. (pp. 183, 188 et 205).

[14] Chrétien, S. and Darses, S., The LASSO for generic design matrices as a function of the relaxation parameter, http://arxiv.org/abs/1105.1430. (pp. 209 et 210).

[15] Donoho, D.L. and Huo, X., Uncertainty principles and ideal atomic decomposition. IEEE Trans. Inform. Theory, 47 (2001) 2845–2862. (p. 177).

[16] Donoho, D.L. and Elad, M.. Optimally sparse representation in general (non-orthogonal) dictionaries via 1 minimization. Proc. Natl. Acad. Sci. USA, 100 (2003) 2197–2202. (p. 177).

[17] Dossal, C., A necessary and sufficient condition for exact recovery by ℓ_1 minimization. (pp. 181, 200, 209 et 212).

[18] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. Least angle regression, Annals of Statistics, 32 (2004) 407–451. (p. 200).

[19] Elad, M. and Bruckstein, A.M., A generalized uncertainty principle and sparse representation in pairs of RN bases. IEEE Trans. Inform. Theory, 48 (2002) 2558–2567. (p. 177).

[20] Fuchs, J.J., On sparse representations in arbitrary redundant bases, IEEE Trans. Inform. Theory, 50 (2004) no. 6 1341–1344. (p. 209).

[21] de la Peña, Victor H. and Giné, E. Decoupling. From dependence to independence. Randomly stopped processes. U-statistics and processes. Martingales and beyond. Probability and its Applications (New York). Springer-Verlag, New York, 1999. (p. 222).

[22] Hiriart-Urruty, J.-B. and Lemaréchal, C. Convex Analysis and Minimization Algorithms II. Advanced theory and bundle methods. Grundlehren der Mathematischen Wissenschaften 306. Springer Verlag. (p. 209).

[23] Kerkyacharian, G.; Mougeot, M.; Picard, D.; Tribouley, K. Learning out of leaders. Multiscale, nonlinear and adaptive approximation, 295–324, Springer, Berlin, 2009. (p. 177).

[24] Koltchinskii, V. Sparse recovery in convex hulls via entropy penalization. Ann. Statist. 37 (2009), no. 3, 1332–1359. (p. 177).

[25] Ledoux, Michel The concentration of measure phenomenon. Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, RI, 2001. (p. 186).

[26] Ledoux, M. and Talagrand, M. Probability in Banach spaces. Isoperimetry and processes. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], 23. Springer-Verlag, Berlin, 1991. xii+480 pp.

[27] Massart, P., Concentration inequalities and model selection. Lectures from the 33rd Summer school on Probability Theory in Saint Flour. Lecture Notes in Mathematics, 1896. Springer Verlag (2007). (p. 208).

[28] Oliveira, R. I., Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. ArXiv:0911.0600, (2010).

[29] Osborne, M. R., Presnell, B. and Turlach, B. A., A new approach to vari- able selection in least squares problems, IMA Journal of Numerical Analysis 20(3) (2000) 389–404. (p. 200).

[30] Osborne, M. R., Presnell, B. and Turlach, B. A., On the LASSO and its dual. J. Comput. Graph. Statist. 9 (2000), no. 2, 319–337. (p. 201).

[31] de la Peña, Victor H., and Montgomery-Smith, S.J. Bounds on the tail probability of U-statistics and quadratic forms. Bull. Amer. Math. Soc. (N.S.) 31 (1994), no. 2, 223–227. (p. 223).

[32] Ralph, D., Global convergence of damped Newton's method for nonsmooth equations via the path search. Math. Oper. Res. 19 (1994), no. 2, 352–389. (p. 201).

[33] Rudelson, M. and Vershynin, R., Geometric approach to error correcting codes and reconstruction of signals. Int. Math. Res. Not. 64 (2005) 4019–4041. (p. 177).

[34] Rudelson, M. and Vershynin, R., Non-asymptotic theory of random matrices: extreme singular values. Proceedings of the International Congress of Mathematicians. Volume III, 1576–1602, Hindustan Book Agency, New Delhi, 2010. (p. 186).

[35] Sun, T. and Zhang C.-H., Comments on: ℓ_1-penalization for mixture regression models, Test (2010) 19, 270–275. (p. 178).

[36] Sun, T. and Zhang, C-H., Scaled sparse linear regression. Biometrika 99 (2012), no. 4, 879–898.

[37] Tao, T., The operator norm of a random matrix. (pp. 38 et 292).

[38] Tibshirani, R. Regression shrinkage and selection via the LASSO, J.R.S.S. Ser. B, 58, no. 1 (1996), 267–288. (pp. 177, 228 et 250).

[39] Tropp, J. A. Norms of random submatrices and sparse approximation. C. R. Math. Acad. Sci. Paris 346 (2008), no. 23-24, 1271–1274. (pp. 181, 187, 193, 218, 219, 222, 282 et 300).

[40] Tropp, J. A. "User friendly tail bounds for sums of random matrices", http://arxiv.org/abs/1004.4389, (2010). (pp. 183, 218 et 226).

[41] van de Geer, S. and Bühlmann, P., On the conditions used to prove oracle results for the Lasso. Electron. J. Stat. 3 (2009) 1360–1392. (pp. 178 et 229).

[42] Wainwright, Martin J., Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic proGramming (Lasso). IEEE Trans. Inform. Theory 55 (2009), no. 5, 2183–2202. (pp. 177, 186, 187, 229 et 250).

[43] Zhao, P. and Yu, B., On model selection consistency of Lasso. J. Mach. Learn. Res. 7 (2006), 2541–2563. (pp. 177 et 229).

Chapter I

Invertibility of random submatrices via tail decoupling and a Matrix Chernoff Inequality

with Sébastien Darses.

Abstract

Let X be a $n \times p$ matrix. We give a new proof of the quasi-isometry property for random submatrices of X obtained by uniform column sampling. The result exhibits explicit constants and some of them are improved by a factor 100. The analysis relies on a tail decoupling argument, of independent interest, and a recent version of the Non-Commutative Chernoff inequality (NCCI).

1 Introduction

Problem statement

Let X be a matrix in $\mathbb{R}^{n\times p}$. The goal of this paper is to propose a new upper bound for the probability that the submatrix X_T fails to be an r_0-quasi isometry when T is a random index subset of size s of $\{1,\dots,p\}$ drawn uniformly at random and X_T is the matrix obtained by extracting the columns of X indexed by T. By an r_0-quasi isometry, we simply mean $\|X_T^t X_T - I\| \leq r_0$. In the sequel, we assume that the columns of X have unit norm.

Proving that the quasi isometry property hold with high probability has applications in Compressed Sensing and high dimensional statistics based on sparsity. The uniform version of the quasi-isometry property, i.e. satisfied for all possible T's, is called the Restricted Isometry Property (RIP) and has been widely studied for random i.i.d. subgaussian matrices [7]. Recent works such as [14] proved that the quasi isometry property holds with high probability for matrices satisfying an certain incoherence assumption. Checking that a matrix is sufficiently incoherent is easy to check in practice. Such types of result are therefore of great potential interest for a wide class of problems involving high dimensional linear or nonlinear regression models.

In a recent work based on the landmark papers of Bourgain and Tzafriri [1] (see also [3]) and Rudelson [8], Tropp proved the following theorem.

Theorem 1.1 *[37, Theorem 1.1] Let A be an $n \times n$ Hermitian matrix, decomposed into diagonal and off-diagonal parts: $A = D+H$. Fix p in $[2, +\infty)$, and set $q = \max\{p, 2\log(n)\}$. Then*

$$\mathbb{E}_p \|RAR\| \leq C \left[q\mathbb{E}_p \|RHR\|_{\max} + \sqrt{\delta q}\mathbb{E}_p \|HR\|_{1,2} + \delta \|H\| \right] + \mathbb{E}_p \|RDR\| .$$

Here, R denotes the square diagonal "selector" matrix whose j^{th} diagonal entry is δ_j, where $\{\delta_j\}$ denotes a sequence of independent Bernoulli 0–1 random variables with common expectation δ, and the symbol $\mathbb{E}_p$ denotes the L_p norm $(\mathbb{E}|\cdot|^p)^{1/p}$. The proof heavily relies on the Non-Commutative Kintchin inequality in a similar way as in [9].

Using this result and Markov's inequality, Candès and Plan proved in [14, Theorem 3.2] that the 1/2-quasi isometry property holds with probability greater than $1 - p^{-2\log(2)}$ when $s \leq p/(4\|X\|^2)$ and the coherence of X, i.e. $\max |X_k^t X_l|$, $k \neq l$, is sufficiently small. The r_0-quasi isometry property then holds with high probability under easily checkable assumptions on X.

Our contribution

The present paper aims at giving a more precise and self-contained version of Theorem 3.2 in [14]. Our result yields explicit constants and some of them are improved by a factor 100. The analysis relies on a tail decoupling argument, of independent interest, and a recent version of the Non-Commutative Chernoff inequality (NCCI) [11].

Additional notations

For $T \subset \{1, \ldots, p\}$, we denote by $|T|$ the cardinal of T. Given a vector $x \in \mathbb{R}^p$, we set $x_T = (x_j)_{j \in T} \in \mathbb{R}^{|T|}$. The canonical scalar product in $\mathbb{R}^p$ is denoted by $\langle \cdot, \cdot \rangle$.

For any matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we denote by A^t its transpose. The set of symmetric real matrices is denoted by $\mathbb{S}_n$. We denote by $\|A\|$ the operator norm of A; $\|A\|_{1\to 2}$ denotes the maximum l_2-norm of a column of A and $\|A\|_{\max}$ is the maximum absolute entry of A. We use the Loewner ordering on symmetric real matrices: if $A \in \mathbb{S}_n$, $0 \preceq A$ is equivalent to saying that A is positive semi-definite, and $A \preceq B$ stands for $0 \preceq B - A$.

The coherence of X, denoted by $\mu(X)$, is defined by

$$\mu(X) \quad = \quad \max_{1 \leq k < l \leq p} |\langle X_k, X_l \rangle|. \tag{1.0}$$

As in [37], we consider the 'hollow Gram' matrix H:

$$H \quad = \quad X^t X - I. \tag{1.1}$$

Recall that R is the diagonal matrix composed of iid Bernoulli variables δ_j, $j = 1, \ldots, p$ with expectation denoted by δ. In the sequel, R' will always denote an independent copy of R. Let R_s be a diagonal matrix whose diagonal is a random vector $\delta^{(s)}$ of length p, uniformly distributed on the set of all vectors with s components equal to 1 and $p - s$ components equal to 0. Notice that when $\delta = s/p$, the support of the diagonal of R has cardinal close to s with high probability, by a standard concentration argument.

2 Main results

Singular value concentration theorem

Theorem 2.1 *Let $r \in (0,1)$, $\alpha \geq 1$. Let us be given a full rank matrix $X \in \mathbb{R}^{n\times p}$ and a positive integer s, such that*

$$\mu(X) \leq \frac{r}{(1+\alpha)\log p} \tag{2.2}$$

$$s \leq \frac{r^2}{(1+\alpha)e^2}\,\frac{p}{\|X\|^2 \log p}. \tag{2.3}$$

Let $T \subset \{1,\dots,p\}$ be a random support with uniform distribution on index sets with cardinal s. Then the following bound holds:

$$\mathbb{P}\left(\|X_T^t X_T - I_s\| \geq r\right) \leq \frac{1944}{p^\alpha}. \tag{2.4}$$

About the various constants

The constant 1944 stems from the following decomposition: 2 (poissonization) $\times$324 (decoupling) $\times$3 (union bound). This constant might look large. However, in many statistical applications as in sparse models, p is often assumed to be very large.

Let us now compare the constants C_s and C_μ in the inequalities

$$\mu(X) \leq \frac{C_\mu}{\log p} \tag{2.5}$$

$$s \leq C_s\,\frac{p}{\|X\|^2 \log p}, \tag{2.6}$$

to the one of [14]. The larger C_s and C_μ are, the better the result is.

One of the various constraints on the rate α in [14] is given by the theorem of Tropp in [37]. In this setting, $\alpha = 2\log 2$ and $r_0 = 1/2$, the author's choice of $1/2$ being unessential. To obtain such a rate α, they need to impose the r.h.s. of (3.15) in [14] to be less than $1/4$, that is $30C_\mu + 13\sqrt{2C_s} \leq \frac{1}{4}$. This yields $C_s < 1.19\times 10^{-4}$. Choosing C_s close to 1.19×10^{-4}, e.g. $C_s \simeq 1.18\ 10^{-4}$, we obtain:

$$C_s \simeq 1.18\ 10^{-4}, \quad C_\mu \simeq 1.7\ 10^{-3}.$$

Our theorem allows to choose any rate $\alpha > 0$. To make a fair comparison, let us choose $\alpha = 2\log 2$ and $r = 1/2$. We obtain:

$$C_s \simeq 0.014, \quad C_\mu = 0.2.$$

3 Proof of Theorem 2.1

In order to study the invertibility condition, we want to obtain bounds for the distribution tail of random sub-matrices of $H = X^t X - I$.

Let R' be an independent copy of R. Let us recall two basic estimates:

$$\|H\|_{1\to 2}^2 \leq \|X\|^2, \quad \|H\|^2 \leq \|X\|^4.$$

As a preliminary, let us notice that

$$\mathbb{P}\left(\|R_s H R_s\| \geq r\right) \;\leq\; 2\,\mathbb{P}\left(\|RHR\| \geq r\right), \tag{3.3}$$

which can be actually proven using the same kind of 'Poissonization argument' as in Claim (3.29) p.2173 in [14].

To study the distribution tail of $\|RHR\|$, we use a decoupling technique which consists of replacing $\|RHR\|$ with $\|RHR'\|$.

Proposition 3.1 *The operator norm of RHR satisfies*

$$\mathbb{P}\left(\|RHR\| \geq r\right) \quad\leq\quad 60\;\mathbb{P}\left(\|RHR'\| \geq r/2\right). \tag{3.4}$$

The main feature of this inequality is that the decoupling constant sits in front of the probability instead of affecting the deviation. In addition to this decoupling argument, we need the following technical concentration result.

Proposition 3.2 *Let $X \in \mathbb{R}^{n\times p}$ be a full rank matrix. For all 4-tuples (s,r,u,v) of parameters such that $\frac{p}{s}\frac{r^2}{e} \geq u^2 \geq \frac{s}{p}\|X\|^4$ and $v^2 \geq \frac{s}{p}\|X\|^2$, the following bound holds:*

$$\mathbb{P}\left(\|RHR'\| \geq r\right) \quad\leq\quad 3\;p\;\mathcal{V}(s,[r,u,v]), \tag{3.5}$$

with

$$\mathcal{V}(s,[r,u,v]) \quad=\quad \left(e\frac{s}{p}\frac{u^2}{r^2}\right)^{\frac{r^2}{v^2}} + \left(e\frac{s}{p}\frac{\|X\|^4}{u^2}\right)^{u^2/\|X\|^2} + \left(e\frac{s}{p}\frac{\|X\|^2}{v^2}\right)^{v^2/\mu(X)^2}.$$

We now have to analyze carefully the various quantities in Proposition 3.2 in order to obtain for $P\left(\|RHR'\| \geq r\right)$ a bound of the order $e^{-\alpha\log p}$.

Set $\alpha' = \alpha + 1$. We tune the parameters so that

$$\frac{u^2}{\|X\|^2} \quad=\quad \alpha'\log p \tag{3.5}$$

$$\frac{v^2}{\mu(X)^2} \quad=\quad \alpha'\log p \tag{3.6}$$

$$\frac{r^2}{v^2} \quad\geq\quad \alpha'\log p, \tag{3.7}$$

and

$$e\frac{s}{p}\frac{\|X\|^4}{u^2} \quad\leq\quad e^{-1} \tag{3.8}$$

$$e\frac{s}{p}\frac{\|X\|^2}{v^2} \quad\leq\quad e^{-1} \tag{3.9}$$

$$e\frac{s}{p}\frac{u^2}{r^2} \quad\leq\quad e^{-1}. \tag{3.10}$$

A crucial quantity turns out to be $\frac{s}{p}\|X\|^2$. Keeping in mind that the hypothesis on the coherence reads

$$\mu(X) \quad\leq\quad \frac{C_\mu}{\log p}, \tag{3.11}$$

it is relevant to impose that s satisfies

$$\frac{s}{p}\|X\|^2 \quad = \quad \frac{C_s}{\log p}, \tag{3.12}$$

where the constants C_μ and C_s will be tuned according to several constraints. The equalities (3.5-3.6) determine the values of u and v. It remains to show that the previous inequalities are satisfied for a suitable choice of C_μ and C_s.

First, plugging (3.5) into (3.10), we obtain:

$$\alpha' \frac{s}{p}\|X\|^2 \log p \quad \leq e^{-2} r^2.$$

Using (3.12), it follows that

$$C_s \quad \leq \quad \frac{r^2}{\alpha' e^2}.$$

Now, the bound (3.8) is satisfied if

$$\frac{e^2 C_s}{\log p} \quad \leq \quad \alpha' \log p.$$

Based on (3.9), it suffices to have $\frac{r^2}{\alpha'^2} \leq \log^2 p$, that is $p \geq e > e^{r/\alpha'}$.

Second, plugging (3.6) into (3.9), we obtain:

$$e^2 \frac{s}{p}\|X\|^2 \quad \leq \quad \alpha' \mu(X)^2 \log p.$$

Using (3.11) and (3.12), it follows that

$$e\sqrt{\frac{C_s}{\alpha'}} \quad \leq \quad C_\mu.$$

Finally, (3.6-3.7) yields $r^2 \geq \alpha'^2 \mu(X)^2 \log^2 p$. In view of (3.11), it thus suffices to have $r \geq \alpha' \; C_\mu$.

As a conclusion, in order to ensure the six previous constraints, it suffices to choose C_s and C_μ such that:

$$C_\mu \leq \frac{r}{1+\alpha} \quad \textit{and} \quad C_s \leq \min\left(\frac{r^2}{(1+\alpha)e^2}, (1+\alpha)\frac{C_\mu^2}{e^2}\right).$$

This completes the proof of Theorem 2.1.

4 Proof of the tail decoupling and the concentration result

Proof of Proposition 3.1

Let us write

$$RHR \quad = \quad \sum_{j \neq k} \delta_j \delta_k H_{jk}.$$

Let $\{\eta_i\}$ be a sequence of iid Rademacher independent of $\mathcal{D} := \{\delta_i, 1 \le i \le p\}$. Following Bourgain and Tzafriri [1], and de la Peña and Giné [6], we build up from these two methods an auxiliary r.v. for our purpose:

$$Z = Z(\eta, \delta) := \sum_{j \neq k} (1 - \eta_j \eta_k) \delta_j \delta_k H_{jk}.$$

We can thus write

$$Z \quad = \quad RHR + Y, \tag{4.5}$$

where

$$Y = \sum_{1 \le i \neq j \le p} B_{ij} \eta_i \eta_j, \quad B_{ij} = B_{ij}(\delta) \in \mathcal{M}_p(\mathbb{R}). \tag{4.6}$$

For the sake of completeness, we recall basic arguments from Corollary 3.3.8 p.12 in de la Peña and Giné [6] (applied to (4.5)) to obtain a lower bound for $\mathbb{P}(\|Z\| \ge \|RHR\|)$. (We henceforth work conditionally on $\mathcal{D}$.)

Hahn-Banach's theorem gives a linear form x^* on $\mathcal{M}_p(\mathbb{R})$ such that

$$\begin{aligned} \mathbb{P}(\|Z\| \ge \|RHR\|) &\ge \mathbb{P}(x^*(Z) \ge x^*(RHR)) \\ &\ge \mathbb{P}(x^*(Y) \ge 0). \end{aligned}$$

Set $\xi := x^*(Y)$. Using Holder's inequality twice, first writing $\mathbb{E}|\xi| = 2\mathbb{E}\ \xi \mathbb{I}_{\xi > 0}$ (since ξ is centered), second noting $\xi^2 = \xi^{2/3} \xi^{4/3}$, one obtains:

$$\mathbb{P}(\xi \ge 0) \ge \frac{1}{4} \frac{(\mathbb{E}|\xi|)^2}{\mathbb{E}\ \xi^2} \ge \frac{1}{4} \frac{(\mathbb{E}\ \xi^2)^2}{\mathbb{E}\ \xi^4}.$$

Hence,

$$\mathbb{P}\left(\|Z\| \ge \|RHR\| \ |\mathcal{D}\right) \quad \ge \quad \frac{1}{4 \times 15} = \frac{1}{60}. \tag{4.4}$$

Multiplying both sides by $\mathbb{I}_{\{\|RHR\| \ge r\}}$ and taking the expectation, one has

$$\frac{1}{60}\ \mathbb{P}(\|RHR\| \ge r) \quad \le \quad \mathbb{P}\left(\|Z\| \ge r\right). \tag{4.5}$$

As from now, we can use the same kind of arguments as in [37, Prop. 2.1]. There is a $\eta^* \in \{-1, 1\}^p$ for which

$$\mathbb{P}\left(\|Z\| \ge r\right) = \mathbb{E}\mathbb{E}\left[\mathbb{I}_{\{\|Z\| \ge r\}} | (\eta_i)\right] \quad \le \quad \mathbb{E}\ \mathbb{I}_{\{\|Z(\eta^*, \delta)\| \ge r\}} = \mathbb{P}(\|Z(\eta^*, \delta)\| \ge r).$$

Hence, setting $T = \{i, \eta_i^* = 1\}$, we can write

$$Z(\eta^*, \delta) \quad = \quad 2 \sum_{j \in T,\ k \in T^c} \delta_j \delta_k H_{jk} + 2 \sum_{j \in T^c,\ k \in T} \delta_j \delta_k H_{jk}. \tag{4.5}$$

Since H is hermitian, we have

$$\left\| \sum_{j \in T,\ k \in T^c} \delta_j \delta_k H_{jk} + \sum_{j \in T^c,\ k \in T} \delta_j \delta_k H_{jk} \right\| = \left\| \sum_{j \in T,\ k \in T^c} \delta_j \delta_k H_{jk} \right\|. \tag{4.5}$$

Now, let (δ_i') be an independent copy of (δ_i). Set $\widetilde{\delta}_i = \delta_i$ if $i \in T$ and $\widetilde{\delta}_i = \delta_i'$ if $i \in T^c$. Since the vectors (δ_i) and $(\widetilde{\delta}_i)$ have the same law, we then obtain:

$$\mathbb{P}\left(\|Z\| \geq r\right) \quad \leq \quad \mathbb{P}\left(2 \left\| \sum_{j\in T,\ k\in T^c} \delta_j \delta_k' H_{jk} \right\| \geq r\right).$$

Re-introducing the missing entries in H yields

$$\mathbb{P}\left(\|Z\| \geq r\right) \quad \leq \quad \mathbb{P}(\|RHR'\| \geq r/2),$$

which concludes the proof of the lemma due to (4.5).

Remark 4.1 *The previous result can be seen as a special case of Theorem 1 p.224 of the seminal paper [5]. Tracking back the various constants involved in this theorem, we obtained the inequality*

$$\mathbb{P}\left(\|RHR\| \geq r\right) \quad \leq \quad 10^3\ \mathbb{P}\left(\|RHR'\| \geq \frac{r}{18}\right). \tag{4.4}$$

Proof of Proposition 3.2

We first apply the NCCI to $\|RHR'\|$ by conditioning on R.

Lemma 4.2 *The following bound holds:*

$$\begin{aligned} P\left(\|RHR'\| \geq r\right) \quad \leq \quad & \mathbb{P}\left(\|RH\| \geq u\right) + \mathbb{P}\left(\|RH\|_{1\to 2} \geq v\right) \\ & + p\left(e\frac{s}{p}\ \frac{u^2}{r^2}\right)^{\frac{r^2}{v^2}}. \end{aligned} \tag{4.4}$$

extbfProof.
We have $\|RHR'\|^2 = \|RHR'^2HR\|$. But $R'^2 = R'$, so

$$\mathbb{P}\left(\|RHR'\| \geq r\right) \quad = \quad P\left(\|RHR'HR\| \geq r^2\right). \tag{4.5}$$

We will first compute the conditional probability

$$\mathbb{P}\left(\|RHR'HR\| \geq r^2 \mid R\right) \quad := \quad \mathbb{E}\left[\mathbb{I}_{\{\|RHR'HR\| \geq r^2\}} \mid R\right]. \tag{4.6}$$

Notice that

$$RHR'HR \ = \ \sum_{j=1}^{p} \delta_j' Z_j Z_j^t \ := \ \sum_{j=1}^{p} A_j.$$

where Z_j is the j^{th} column of RH, $j = 1, \ldots, p$.
Since $\sum_{j=1}^{p} Z_j Z_j^t = RH^2R$ and $\|Z_j Z_j^t\| = \|Z_j\|_2^2$, we then obtain

$$\|A_j\| \quad \leq \quad \|RH\|_{1\to 2}^2 \tag{4.6}$$

$$\left\| \sum_{j=1}^{p} \mathbb{E}A_j \right\| \quad \leq \quad \frac{s}{p}\|RH\|^2. \tag{4.7}$$

The NCCI then yields

$$\mathbb{P}\left(\|RHR'HR\| \geq r^2 \mid R\right) \quad \leq \quad p\left(e\frac{s}{p}\frac{\|RH\|^2}{r^2}\right)^{r^2/\|RH\|_{1\to 2}^2}, \tag{4.8}$$

provided that

$$e\frac{s}{p}\frac{\|RH\|^2}{r^2} \quad \leq \quad 1. \tag{4.9}$$

Let us now introduce the events

$$\mathcal{A} = \left\{\|RHR'HR\| \geq r^2\right\}; \quad \mathcal{B} = \{\|RH\| \geq u\}; \quad \mathcal{C} = \{\|RH\|_{1\to 2} \geq v\}.$$

We have

$$\begin{aligned}\mathbb{P}(\mathcal{A}) &= \mathbb{P}(\mathcal{A} \mid \mathcal{B}\cup\mathcal{C})\mathbb{P}(\mathcal{B}\cup\mathcal{C}) + \mathbb{P}(\mathcal{A}\cap\mathcal{B}^c\cap\mathcal{C}^c)\\ &\leq \mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{C}) + \mathbb{P}(\mathcal{A}\cap\mathcal{B}^c\cap\mathcal{C}^c).\end{aligned}$$

The identity $\mathbb{P}(\mathcal{A}\cap\mathcal{B}^c\cap\mathcal{C}^c) = \mathbb{E}\left[\mathbb{I}_{\mathcal{A}\cap\mathcal{B}^c\cap\mathcal{C}^c}\right] = \mathbb{E}\left[\mathbb{P}\left(\mathcal{A} \mid R\right)\, \mathbb{I}_{\mathcal{B}^c\cap\mathcal{C}^c}\right]$ concludes the lemma. □

We now have to control the norm of $\frac{s}{p}RH^2R$, the norm of RH and the column norm of RH. Let us begin with $\|RH\| = \|HR\|$.

Lemma 4.3 *The following bounds hold:*

$$\begin{aligned}P\left(\|HR\| > u\right) &\leq p\left(e\frac{s}{p}\frac{\|X\|^4}{u^2}\right)^{u^2/\|X\|^2}\\ \mathbb{P}\left(\|RH\|_{1\to 2} \geq v\right) &\leq p\left(e\frac{s}{p}\frac{\|X\|^2}{v^2}\right)^{v^2/\mu(X)^2},\end{aligned}$$

provided that $e\frac{s}{p}\frac{\|X\|^4}{u^2}$ *and* $e\frac{s}{p}\frac{\|X\|^2}{v^2}$ *are less than* 1.

extbfProof.

The steps are of course the same as what we have just done in the proof of Lemma 3.1. Notice that

$$\mathbb{P}\left(\|RH\| > u\right) = \mathbb{P}\left(\|HR\|^2 > u^2\right) = \mathbb{P}\left(\|HRH\| > u^2\right).$$

The j^{th} column of H is $H_j = X^tX_j - e_j$. Moreover,

$$HRH = \sum_{j=1}^{p}\delta_j H_j H_j^t. \tag{4.4}$$

We have $\|H_jH_j^t\| = \|H_j\|_2^2 \leq \|H\|_{1\to 2}^2 \leq \|X\|^2$, and

$$\left\|\sum_{j=1}^{p}\mathbb{E}[\delta_j H_j H_j^t]\right\| \quad \leq \quad \frac{s}{p}\|H\|^2 \leq \frac{s}{p}\|X\|^4. \tag{4.5}$$

We finally deduce from the NCCI that

$$\mathbb{P}\left(\|HRH\| \geq u^2\right) \quad \leq \quad p\left(e\frac{s}{p}\frac{\|X\|^4}{u^2}\right)^{u^2/\|X\|^2}. \tag{4.6}$$

Let us now control the supremum ℓ_2-norm of the columns of RH. Set

$$M \quad = \quad \sum_{k=1}^{p} \delta_k \; diag(H_k H_k^t). \tag{4.7}$$

Notice that

$$\|RH\|_{1\to 2}^2 \quad = \max_{k=1}^{p} \|(RH)_k\|_2^2 = \|diag\left((RH)^t RH\right)\| = \quad \left\|diag\left(H^t RH\right)\right\|.$$

Thus,

$$\|RH\|_{1\to 2}^2 \quad = \quad \left\| diag\left(\sum_{k=1}^{p} \delta_k (H^t)_k H_k^t \right) \right\|.$$

Using symmetry of H and interchanging the summation and the diag, we obtain that $\|RH\|_{1\to 2}^2 = \|M\|$. Moreover, we have for all $k \in \{1, \cdots, p\}$,

$$\|diag(H_k H_k^t)\| \quad = \max_{j=1}^{p} (X_j X_k)^2 \leq \quad \mu(X)^2, \tag{4.6}$$

and

$$\|\mathbb{E}M\| = \frac{s}{p}\|diag(HH^t)\|^2 = \frac{s}{p}\|H\|_{1\to 2}^2 \leq \frac{s}{p}\|X\|^2.$$

Applying the NCCI completes the lemma. □ This lemma concludes the proof of Proposition 3.2.

5 Appendix

On Rademacher chaos of order 2

Lemma 5.1 *Let ξ be an homogeneous Rademacher chaos of order 2. Then*

$$\mathbb{E}\ \xi^4 \quad \leq \quad 15\ (\mathbb{E}\ \xi^2)^2. \tag{5.6}$$

extbfProof. We develop both ξ^2 and ξ^4 (as Littlewood's proof of Kintchine's inequality). The multinomial formula applied to the chaos $\xi = \sum_{i<j} x_{ij}\eta_i\eta_j$ with an integer q, gives

$$\xi^q \quad = \quad \sum \frac{q!}{\prod \alpha_{ij}!} \prod x_{ij}^{\alpha_{ij}} (\eta_i\eta_j)^{\alpha_{ij}}, \tag{5.7}$$

where the sum is over all the α_{ij}'s, $i < j$, such that $\sum_{i<j} \alpha_{ij} = q$, and the products are over all the index (i, j), $i < j$, ordered via the lexicographical order. Hereafter, we adopt these conventions when considering any product of the α_{ij}'s and the x_{ij}'s.

Case $q = 2$ — The partitions of 2 are $2 + 0's$ and $1 + 1 + 0's$. Consider the partition $1+1+0's$, say $\alpha_{kl} = \alpha_{k'l'} = 1$ for some 4-uple (k, l, k', l') with $k \leq k'$. We have $(k, l) \neq (k', l')$, $k < l$ and $k' < l'$. Thus,

$$\mathbb{E}[\eta_k\eta_l\eta_{k'}\eta_{l'}] = \begin{cases} \mathbb{E}[\eta_k]\mathbb{E}[\eta_l\eta_{k'}\eta_{l'}](= 0) \text{ if } k < k' \\ \\ \mathbb{E}[\eta_k^2]\mathbb{E}[\eta_l]\mathbb{E}[\eta_{l'}](= 0) \text{ esle.} \end{cases}$$

Therefore, $\mathbb{E}\,\xi^2$ only depends on the partition $2+0's$, and one has

$$\mathbb{E}\,\xi^2 \quad = \quad \sum_{i<j} x_{ij}^2.$$

Case $q=4$ — The partitions of 4 are 4, $2+2$, $3+1$, $2+1+1$ and $1+1+1+1$ (we now omit the zeros).

First, using the same arguments as in the case $q=2$, we show that the terms in $\mathbb{E}\,\xi^4$ corresponding to the partitions $3+1$ and $2+1+1$ vanish.

Second, the partitions $1+1+1+1$ involve four different couples (i,i'), (j,j'), (k,k') and (l,l') (recall that $i<i'$, etc., and that the couples are lexicographically ordered). The only terms corresponding to the partitions $1+1+1+1$ whose expectation does not vanish are of the form

$$x_{i_1i'_1}x_{i_1i'_2}x_{i_2i'_2}x_{i_2i'_1}\ \eta_{i_1}^2\eta_{i'_1}^2\eta_{i_2}^2\eta_{i'_2}^2 = x_{i_1i'_1}x_{i_1i'_2}x_{i_2i'_2}x_{i_2i'_1}.$$

Finally, the α_{ij}'s corresponding to the partitions 4 and $2+2$ are even: $\alpha_{ij}=2\beta_{ij}$ where $\sum\beta_{ij}=2$. Therefore

$$\begin{aligned}\mathbb{E}\,\xi^4 \quad &= \quad \sum\frac{4!}{\prod(2\beta_{ij})!}\prod x_{ij}^{2\beta_{ij}} + \sum 4!\ x_{i_1i'_1}x_{i_1i'_2}x_{i_2i'_2}x_{i_2i'_1}\\ &\leq \quad 3\sum\frac{2!}{\prod\beta_{ij}!}\prod(x_{ij}^2)^{\beta_{ij}} + \frac{4!}{2}\sum\left(x_{i_1i'_1}^2x_{i_2i'_2}^2 + x_{i_1i'_2}^2x_{i_2i'_1}^2\right).\end{aligned}$$

But

$$\begin{aligned}\sum\frac{2!}{\prod\beta_{ij}!}\prod(x_{ij}^2)^{\beta_{ij}} \quad &= \quad \left(\sum_{i<i'}x_{ii'}^2\right)^2\\ \sum_{\text{rectangles dans partie superieure}} 2\left(x_{i_1i'_1}^2x_{i_2i'_2}^2 + x_{i_1i'_2}^2x_{i_2i'_1}^2\right) \quad &\leq \quad \sum 2!\ x_{ii'}x_{jj'}\\ &= \quad \left(\sum_{i<i'}x_{ii'}^2\right)^2.\end{aligned}$$

Hence,

$$\mathbb{E}\,\xi^4 \leq \quad (3+6)\ (\mathbb{E}\,\xi^2)^2.$$

A Non-Commutative Chernoff inequality

We need a corollary of a Matrix Chernoff's inequality recently established in [11].

Theorem 5.2 *(Matrix Chernoff) Let $X_1,\ldots,X_p$ be independent random positive semi-definite matrices taking values in $\mathbb{R}^{d\times d}$. Set $S_p=\sum_{j=1}^p X_j$. Assume that for all $j\in\{1,\ldots,p\}$ $\|X_j\|\leq B$ a.s. and*

$$\|S_p\| \quad \leq \quad \mu_{\max}.$$

Then, for all $t\geq e\ \mu_{\max}$,

$$\mathbb{P}\left(\|S_p\|\geq r\right) \quad \leq \quad d\exp\left(\frac{e\ \mu_{\max}}{r}\right)^{r/B}.$$

(Set $r=(1+\delta)\mu_{\max}$ and use $e^\delta\leq e^{1+\delta}$ in Theorem 1.1 [11].)

Bibliography

[1] Bourgain, J., Tzafriri, L., Invertibility of "large" submatrices with applications to the geometry of Banach spaces and harmonic analysis. Israel J. Math. 57 (1987), no. 2, 137–224. (pp. 217 et 222).

[2] Candès, E. J. and Plan, Y. Near-ideal model selection by ℓ_1 minimization. Ann. Statist. 37 (2009), no. 5A, 285–2177. (pp. 177, 178, 180, 181, 182, 183, 184, 185, 187, 188, 189, 192, 193, 205, 206, 207, 217, 218, 219, 220, 228, 229, 230, 233, 243, 245, 250, 254, 255, 256, 286, 300 et 301).

[3] Ledoux, M. and Talagrand, M. Probability in Banach spaces. Isoperimetry and processes. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], 23. Springer-Verlag, Berlin, 1991. xii+480 pp. (p. 217).

[4] Oliveira, R.I. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. ArXiv:0911.0600, (2009).

[5] de la Peña, V.H., and Montgomery-Smith, S.J. Bounds on the tail probability of U-statistics and quadratic forms. Bull. Amer. Math. Soc. (N.S.) 31 (1994), no. 2, 223–227. (p. 223).

[6] de la Peña, V.H. and Giné, E. Decoupling. From dependence to independence. Randomly stopped processes. U-statistics and processes. Martingales and beyond. Probability and its Applications (New York). Springer-Verlag, New York, 1999. (p. 222).

[7] Mendelson, S. , Pajor A. and Tomczak-Jaegermann N. Uniform uncertainty principle for Bernoulli and subgaussian ensembles, Constructive Approximation, 28 (2008), no. 3, 277-289. (p. 217).

[8] Rudelson M. Random vectors in isotropipc position, J. Funct. Anal. 164 (1999), no. 1, 60–72. (p. 217).

[9] Rudelson, M and Vershynin, R. Sampling from large matrices: an approach through geometric functional analysis, Journal of the ACM (2007). (p. 218).

[10] Tropp, J. A. Norms of random submatrices and sparse approximation. C. R. Math. Acad. Sci. Paris 346 (2008), no. 23-24, 1271–1274. (pp. 181, 187, 193, 218, 219, 222, 282 et 300).

[11] Tropp, J. A. User friendly tail bounds for sums of random matrices. Found. Comput. Math., (2011).
(pp. 183, 218 et 226).

Chapter J

On prediction with the LASSO when the design is not incoherent

Abstract

The LASSO estimator is an ℓ_1-norm penalized least-squares estimator, which was introduced for variable selection in the linear model. When the design matrix satisfies, e.g. the Restricted Isometry Property, or has a small coherence index, the LASSO estimator has been proved to recover, with high probability, the support and sign pattern of sufficiently sparse regression vectors. Under similar assumptions, the LASSO satisfies adaptive prediction bounds in various norms. The present note provides a prediction bound based on a new index for measuring how favorable is a design matrix for the LASSO estimator. We study the behavior of our new index for matrices with independent random columns uniformly drawn on the unit sphere. Using the simple trick of appending such a random matrix (with the right number of columns) to a given design matrix, we show that a prediction bound similar to [14, Theorem 2.1] holds without any constraint on the design matrix, other than restricted non-singularity.

Keywords: LASSO; Coherence; Restricted Isometry Property; ℓ_1-penalization; High dimensional linear model.

1 Introduction

Given a linear model

$$y = X\beta + \varepsilon \tag{1.1}$$

where $X \in \mathbb{R}^{n\times p}$ and ε is a random vector with gaussian distribution $\mathcal{N}(0, \sigma^2 I)$ the LASSO estimator is given by

$$\hat{\beta} \quad = \quad argmin_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1. \tag{1.2}$$

This estimator was first proposed in the paper of Tibshirani [23]. The LASSO estimator $\hat{\beta}$ is often used in the high dimensional setting where p is much larger than n. As can be expected, when $p \gg n$, estimation of β is hopeless in general unless some additional property of β is assumed. In many practical situations, it is considered relevant to assume that β is sparse, i.e. has only a few nonzero components, or at least compressible, i.e. the

magnitude of the non zero coefficients decays with high rate. It is now well recognized that the ℓ_1 penalization of the likelihood often promotes sparsity under certain assumptions on the matrix X. We refer the reader to the book [4] and the references therein for a state of the art presentation of the LASSO and the tools involved in the theoretical analysis of its properties. One of the main interesting properties of the LASSO estimator is that it is a solution of a convex optimization problem and it can be computed in polynomial time, i.e. very quickly in the sense of computational complexity theory. This makes a big difference with other approaches based on variable selection criteria like AIC [1], BIC [17], Foster and George's Risk Inflation Criterion [12], etc, which are based on enumeration of the possible models, or even with the recent proposals of Dalalyan, Rigollet and Tsybakov [15], [9], although enumeration is replaced with a practically more efficient Monte Carlo Markov Chain algorithm.

In the problem of estimating $X\beta$, i.e. the prediction problem, it is often believed that the price to pay for reducing the variable selection approach to a convex optimization problem is a certain set of assumptions on the design matrix X *. One of the main contributions of [15] is that no particular assumption on X is required for the prediction problem, as opposed to the known results concerning the LASSO such that [3], [2], [14] and [20], and the many references cited in these works.

An impressive amount of work has been done in the recents years in order to understand the properties of $\hat{\beta}$ unded various assumptions on X. See the recent book by P. Buhlmann and S. Van de Geer [4] for the state of the art. Two well known assumptions on the design matrix are

- small Coherence $\mu(X)$
- small Restricted Isometry Constant $\delta(X)$

where the Coherence $\mu(X)$ is defined as

$$\mu(X) \quad = \quad \max_{j,j'} |X_j^t X_{j'}|,$$

and the Restricted Isometry Constant $\delta(X)$ is the smallest δ such that

$$(1-\delta)\|\beta_T\|_2 \leq \|X_T\beta_T\|_2 \leq (1+\delta)\|\beta_T\|_2 \tag{1.2}$$

for all subset T with cardinal s and all $\beta \in \mathbb{R}^p$. Other conditions are listed in [5]; see Figure 1 in that paper for a diagram summarizing all relationships between them. The Restricted Isometry property is very stringent and implies almost other conditions. Moreover, the Restricted Isometry Constant is NP-hard to compute for general matrices. On the other hand, the Coherence only requires of the order $np(p-1)$ elementary operations. However, it was proved in [14] that a small coherence, say of the order of $1/\log(p)$, is sufficient to prove a property very close to the Restricted Isometry Property: (1.2) holds for a large proportion of subsets $T \subset \{1,\ldots,p\}$, $|T| = s$ (of the order $1 - 1/p^\alpha$, $\alpha > 0$). This result was later refined in [13] with better constants using the recently discovered Non-Commutative deviation inequalities [19]. Less stringent properties are the restricted eigenvalue, the irrepresentable and the compatibility properties.

The goal of this short note is to show that, using a very simple trick, one can prove prediction bounds similar to [14, Theorem 2.1] without any assumption on the design matrix X at the low expense of appending to X a random matrix with independent columns uniformly distributed on the sphere.

*Conditions for model selection consistency are given in e.g. [22], and for exact support and sign pattern recovery with finite samples and $p \gg n$, in [14], [28].

For this purpose, we introduce a new index for design matrices, denoted by $\gamma_{s,\rho_-}(X)$ that allows to obtain novel adaptive bounds on the prediction error. This index is defined for any $s \leq n$ and $\rho_- \in (0,1)$ as

$$\gamma_{s,\rho_-}(X) \quad = \quad \sup_{v \in B(0,1)} \inf_{I \subset \mathcal{S}_{s,\rho_-}} \|X_I^t v\|_\infty, \tag{1.3}$$

where $\mathcal{S}_{s,\rho_-}(X)$ is the family of all S of $\{1,\ldots,p\}$ with cardinal $|S| = s$, such that $\sigma_{\min}(X_S) \geq \rho_-$. The meaning of the index γ_{s,ρ_-} is the following: for any $v \in \mathbb{R}^n$, we look for the "almost orthogonal" family inside the set of columns of X with cardinal s, which is the most orthogonal to v .

One major advantage of this new parameter is that imposing the condition that γ_{s,ρ_-} is small is much less stringent than previous criteria required in the litterature. In particular, many submatrices of X may be very badly conditioned or even singular without altering the smallness of γ_{s,ρ_-}. Computing the new index $\gamma_{s,\rho_-}(X)$ for random matrices with independent columns uniformly distributed on the sphere [†], shows that a prediction bound involving $\gamma_{s,\rho_-}(X)$ can be obtained which is of the same order as the bound of [14, Theorem 2.1].

One very nice property of the index γ_{s,ρ_-} is that it decreases after the operation appending any matrix to a given one. As a very nice consequence of this observation, the results obtained for random matrices can be extended to any matrix X to which a random matrix is appended. This trick can be used to prove new prediction bounds for a modified version of the LASSO obtained by appending a random matrix to any given design matrix. This simple modification of the LASSO retains the fundamental property of being polynomial time solvable unlike the recent approaches based on non-convex criteria for which no computational complexity analysis is available.

The plan of the paper is as follows. In Section 2 we present the index γ_{s,ρ_-} for X and provide an upper bound on this index for random matrices with independent columns uniformly distributed on the sphere, holding with high probability. Then, we present our prediction bound in Theorem 3.1: we give a bound on the prediction squared error $\|X(\beta - \hat{\beta}\|_2^2$ which depends linearly on s. This result is similar in spirit to [14, Theorem 1.2]. The proofs of the above results are given in Section 3. In Section 4, we show how these results can be applied in practice to any problem with a matrix for which γ_{s,ρ_-} is unknown by appending to X an $n \times p_0$ random matrix with i.i.d. columns uniformly distributed on the unit sphere of $\mathbb{R}^n$ and with only a small number p_0 of columns. An appendix contains the proof of some intermediate results.

Notations and preliminary assumptions

A vector β in $\mathbb{R}$ is said to be s-sparse if exactly s of its components are different from zero. Let ρ_- be a positive real number. In the sequel, we will denote by $\mathcal{S}_{s,\rho_-}(X)$ the family of all index subsets S of $\{1,\ldots,p\}$ with cardinal $|S| = s$, such that for all $S \in \mathcal{S}_{s,\rho_-}$, $\sigma_{\min}(X_S) \geq \rho_-$.

[†]or equivalently, post-normalized Gaussian i.i.d. matrices with components following $\mathcal{N}(0, 1/n)$.

2 Main results

A new index for design matrices

Definition 2.1 *The index $\gamma_{s,\rho_-}(X)$ associated with the matrix X in $\mathbb{R}^{n\times p}$ is defined by*

$$\gamma_{s,\rho_-}(X) \quad = \quad \sup_{v\in B(0,1)} \inf_{I\subset \mathcal{S}_{s,\rho_-}} \|X_I^t v\|_\infty. \tag{2.4}$$

An important remark is that the function $X \mapsto \gamma_{s,\rho_-}(X)$ is nonincreasing in the sense that if we set $X'' = [X, X']$, where X' is a matrix in $\mathbb{R}^{n\times p'}$, then $\gamma_{s,\rho_-}(X) \geq \gamma_{s,\rho_-}(X')$.

Unlike the coherence $\mu(X)$, for fixed n and s, the quantity $\gamma_{s,\rho_-}(X)$ is very small for p sufficiently large, at least for random matrices such as normalized standard Gaussian matrices as shown in the following proposition.

Proposition 2.2 *Assume that X is random matrix in $\mathbb{R}^{n\times p}$ with i.i.d. columns with uniform distribution on the unit sphere of $\mathbb{R}^n$. Let ρ_- and $\varepsilon \in (0,1)$, $C_\kappa \in (0,+\infty)$ and $p_0 \in \{\lceil e^{\frac{6}{\sqrt{2\pi}}}\rceil, \ldots, p\}$. Set*

$$K_\varepsilon \quad = \quad \frac{\sqrt{2\pi}}{6}\left((1+C_\kappa)\log\left(1+\frac{2}{\varepsilon}\right)+C_\kappa+\log\left(\frac{C_\kappa}{4}\right)\right).$$

Assume that n, κ and s satisfy

$$n \geq 6, \tag{2.4}$$

$$\kappa = \max\left\{4e^{-2(\ln(2)-1)}, \tfrac{4e^3}{(1-\rho_-)^2}\ \left(\tfrac{(1+K_\varepsilon)(1+C_\kappa)}{c(1-\varepsilon)^4}\right)^2 \log^2(p_0)\log(C_\kappa n)\right\}, \tag{2.4}$$

$$\frac{\max\{\kappa s, 2\times 36\times 3\times 3, \exp((1-\rho_-)/2))}{C_\kappa} \leq n \leq \min\left\{\left(\frac{p_0}{\log(p_0)}\right)^2, \frac{\exp\left(\frac{1-\rho_-}{\sqrt{2}}p_0\right)}{C_\kappa}\right\} \tag{2.4}$$

Then, we have

$$\gamma_{s,\rho_-}(X) \quad \leq \quad 80\ \frac{\log(p_0)}{p_0} \tag{2.5}$$

with probability at least $1 - 5\ \frac{n}{p_0\ \log(p_0)^{n-1}} - 9\ p_0^{-n}$.

Remark 2.3 *Notice that the constraints (2.4) and (2.5) together imply the following constraint on s:*

$$s \leq \quad C_{sparsity}\ \frac{n}{\log^2(p_0)\log(C_\kappa n)}$$

with

$$C_{sparsity} \quad = \quad \frac{c^2(1-\rho_-)^2(1-\varepsilon)^8}{4e^3}\ \frac{C_\kappa}{(1+K_\varepsilon)^2(1+C_\kappa)^2}.$$

A bound of $\|X(\beta - \hat{\beta})\|_2^2$ based on $\gamma_{s,\rho_-}(X)$

In the remainder of this paper, we will assume that the columns of X are ℓ_2-normalized. The main result of this paper is the following theorem.

Theorem 2.4 *Let $\rho_- \in (0,1)$. Let ν be a positive real such that*

$$\nu\ \gamma_{\nu n,\rho_-}(X) \leq \frac{\rho_-\ \sigma_{\min}(X_S)}{n\ \max_{\substack{T\subset\{1,\dots,p\}\\ |T|\leq n}} \sigma_{\max}(X_T)}. \tag{2.4}$$

Assume that $s \leq \nu n$. Assume that β has support S with cardinal s and that

$$\lambda \geq \tag{2.5}$$
$$\sigma\left(B_{X,\nu,\rho_-}\ \max_{\substack{T\subset\{1,\dots,p\}\\ |T|\leq n}} \sigma_{\max}(X_T)\sqrt{2\alpha\ \log(p)+\log(2\nu n)} + \sqrt{(2\alpha+1)\ \log(p)+\log(2)}\right)$$

with

$$B_{X,\nu,\rho_-} = \frac{\nu n\ \gamma_{\nu n,\rho_-}(X)}{\rho_-\ \sigma_{\min}(X_S) - \nu n\ \gamma_{\nu n,\rho_-}(X)\ \max_{\substack{T\subset\{1,\dots,p\}\\ |T|\leq n}} \sigma_{\max}(X_T)}. \tag{2.5}$$

Then, with probability greater than $1-p^{-\alpha}$, we have

$$\frac{1}{2}\|X(\hat{\beta}-\beta)\|_2^2 \leq s\ C_{n,p,\rho_-,\alpha,\nu,\lambda} \tag{2.6}$$

with

$$C_{n,p,\rho_-,\alpha,\nu,\lambda} = \frac{\lambda+\sigma\sqrt{(2\alpha+1)\ \log(p)+\log(2)}}{\rho_-\ \sigma_{\min}(X_S)}\left(\sigma\sqrt{2\alpha\ \log(p)+\log(2\nu n)}+\lambda\right) \tag{2.7}$$

Comments

Equation (2.4) in Theorem 3.1 requires that

$$\gamma_{\nu n,\rho_-}(X) < \rho_- \frac{\sigma_{\min}(X_S)}{\nu n\ \max_{\substack{T\subset\{1,\dots,p\}\\ |T|\leq n}} \sigma_{\max}(X_T)}. \tag{2.8}$$

Proposition 2.2 proves that for random matrices with independent columns uniformly drawn on the unit sphere of $\mathbb{R}^n$ (i.e. normalized i.i.d. gaussian matrices),

$$\gamma_{s,\rho_-}(X) \leq 80\ \frac{\log(p_0)}{p_0} \tag{2.9}$$

with high probability. The case of general design matrices can be treated using a simple trick. It will be studied in Section 4.

The main advantage of using the parameter $\gamma_{\nu n,\rho_-}(X)$ is that it allows X to contain extremely badly conditioned submatrices, a situation that may often occur in practice when certain covariates are very correlated. This is in contrast with the Restricted Isometry Property or the Incoherence condition, or other conditions often required in the litterature. On the other hand, the parameter $\gamma_{\nu n,\rho_-}(X)$ is not easily computable. We will see however in Section 4 how to circumvent this problem in practice by the simple trick consisting of

appending a random matrix with p_0 columns to the matrix X in order to ensure that X satisfies (2.8) with high probability.

Finally, notice that unlike in [14, Theorem 2.1], we make no assumption on the sign pattern of β. In particular, we do not require the sign pattern of the nonzero components to be random. Moreover, the extreme singular values of X_S are not required to be independent of n nor p and the condition (2.8) is satisfied for a wide range of configurations of the various parameters involved in the problem.

3 Proofs

Proof of Proposition 2.2

Constructing an outer approximation for I in the definition of γ_{s,ρ_-}

Take $v \in \mathbb{R}^n$. We construct an outer approximation $\tilde{I}$ of I into which we be able to extract the set I. We procede recursively as follows: until $|\tilde{I}| = \min\{\kappa s, p_0/2\}$, for some positive real number κ to be specified later, do

- Choose $j_1 = argmin_{j=1,\dots,p_0} |\langle X_j, v\rangle|$ and set $\tilde{I} = \{j_1\}$
- Choose $j_2 = argmin_{j=1,\dots,p_0,\ j\notin\tilde{I}} |\langle X_j, v\rangle|$ and set $\tilde{I} = \tilde{I} \cup \{j_2\}$
- $\cdots$
- Choose $j_k = argmin_{j=1,\dots,p_0,\ j\notin\tilde{I}} |\langle X_j, v\rangle|$ and set $\tilde{I} = \tilde{I} \cup \{j_k\}$.

An upper bound on $\|X_{\tilde{I}}^t v\|_\infty$

If we denote by Z_j the quantity $|\langle X_j, v\rangle|$ and by $Z_{(r)}$ the r^{th} order statistic, we get that

$$\|X_{\tilde{I}}^t v\|_\infty \quad = \quad Z_{(\kappa s)}.$$

Since the X_j's are assumed to be i.i.d. with uniform distribution on the unit sphere of $\mathbb{R}^n$, we obtain that the distribution of $Z_{(r)}$ is the distribution of the r^{th} order statistics of the sequence $|X_j^t v|$, $j = 1, \dots, p_0$. By (5) p.147 [13], $|X_j^t v|$ has density g and CDF G given by

$$g(z) \quad = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \left(1 - z^2\right)^{\frac{n-3}{2}} \text{ and } G(z) = \quad 2 \int_0^z g(\zeta)\, d\zeta.$$

Thus,

$$F_{Z_{(r)}}(z) \quad = \quad \mathbb{P}\left(B \geq r\right)$$

where B is a binomial variable $\mathcal{B}\left(p_0, G(z)\right)$. Our next goal is to find the smallest value z_0 of z which satisfies

$$F_{Z_{(\kappa s)}}(z_0) \quad \geq \quad 1 - p_0^{-n}. \tag{3.7}$$

We have the following standard concentration bound for B (e.g. [11]):

$$\mathbb{P}\left(B \leq (1-\varepsilon)\mathbb{E}[B]\right) \quad \leq \quad \exp\left(-\frac{1}{2}\,\varepsilon^2 \mathbb{E}[B]\right)$$

which gives

$$\mathbb{P}\left(B \geq (1-\varepsilon)p_0 G(z)\right) \quad \geq \quad 1 - \exp\left(-\frac{1}{2}\,\varepsilon^2 p_0 G(z)\right)$$

We thus have to look for a root (or at least an upper bound to a root) of the equation

$$G(z) \quad = \quad \frac{1}{\frac{1}{2}\,\varepsilon^2}\,\frac{n}{p_0}\,\log(p_0).$$

Notice that

$$\begin{aligned} G(z) \quad &= \quad 2\,\frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\,\int_0^z \left(1-\zeta^2\right)^{\frac{n-3}{2}}\,d\zeta, \\ &\geq \quad \frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\,z \end{aligned}$$

for $z \leq 1/\sqrt{2}$. By a straightforward application of Stirling's formula (see e.g. (1.4) in [14]), we obtain

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \quad \geq \quad \frac{e^{2\ln(2)}}{2}\,\frac{(n-3)^{3/2}}{(n-2)^{1/2}}.$$

Thus, any choice of z_0 satisfying

$$z_0 \quad \geq \quad \frac{2\,\sqrt{\pi}}{e^{2\ln(2)}}\,\frac{(n-2)^{1/2}}{(n-3)^{3/2}}\,\frac{1}{\frac{1}{2}\,\varepsilon^2}\,\frac{n}{p_0}\,\log(p_0) \tag{3.2}$$

is an upper bound to the quantile for $(1-\varepsilon)p_0 G(z_0)$-order statistics at level p_0^{-n}. We now want to enforce the constraint that

$$(1-\varepsilon)p_0 G(z_0) \quad \leq \quad \kappa s.$$

By again a straightforward application of Stirling's formula, we obtain

$$G(z) \quad \leq \quad \frac{1}{\sqrt{\pi}}\,\frac{e^2}{2}\,\frac{(n-3)^{3/2}}{(n-2)^{1/2}}\,z$$

for $n \geq 4$. Thus, we need to impose that

$$z_0 \quad \leq \quad \frac{2\sqrt{\pi}}{e^2}\,\frac{(n-2)^{1/2}}{(n-3)^{3/2}}\,\frac{\kappa s}{(1-\varepsilon)p_0}. \tag{3.1}$$

Notice that the constraints (3.2) and (3.1) are compatible if

$$\kappa \quad \geq \quad \frac{4}{e^{2(\ln(2)-1)}}\,\frac{1-\varepsilon}{\varepsilon^2}\,\frac{n}{s}\,\log(p_0).$$

Take $\varepsilon = 1 - \frac{1}{n/s\,\log(p_0)}$ and obtain

$$\mathbb{P}\left(\|X_{\tilde{I}}^t v\|_\infty \geq \frac{8\,\sqrt{\pi}}{e^{2\ln(2)}}\,\frac{(n-2)^{1/2}}{(n-3)^{3/2}}\,\frac{n}{p_0}\,\log(p_0)\right) \quad \leq \quad p_0^{-n}$$

for

$$\kappa \quad = \quad \frac{4}{e^{2(\ln(2)-1)}}$$

for any p_0 such that $n/s\log(p_0) \geq \sqrt{2}$, which is clearly the case as soon as $p_0 \geq e^{\frac{6}{\sqrt{2\pi}}}$ for $s \leq n$ as assumed in the proposition.

If $n \geq 6$, we can simplify (3.1) with

$$\mathbb{P}\left(\|X_{\tilde{I}}^t v\|_\infty \geq 80\ \frac{\log(p_0)}{p_0}\right) \quad \leq \quad p_0^{-n} \tag{3.-1}$$

Extracting a well conditionned submatrix of $X_{\tilde{I}}$

The method for extracting X_I from $X_{\tilde{I}}$ uses random column selection. For this purpose, we will need to control the coherence and the norm of $X_{\tilde{I}}$.

Step 1: The coherence of $X_{\tilde{I}}$. Let us define the spherical cap

$$\mathcal{C}(v,h) \quad = \quad \{w \in \mathbb{R}^n \mid \langle v, w\rangle \geq h\}.$$

The area of $\mathcal{C}(v,h)$ is given by

$$Area\left(\mathcal{C}(v,h)\right) \quad = \quad Area(\mathcal{S}(0,1)) \int_0^{2h-h^2} t^{\frac{n-1}{2}}(1-t)^{\frac{1}{2}}dt.$$

Thus, the probability that a random vector w with Haar measure on the unit sphere $\mathcal{S}(0,1)$ falls into the spherical cap $\mathcal{C}(v,h)$ is given by

$$\begin{aligned}\mathbb{P}\left(w \in \mathcal{C}(v,h)\right) &= \frac{\mathcal{C}(v,h)}{\mathcal{S}(0,1)} \\ &= \frac{\int_0^{2h-h^2} t^{\frac{n-1}{2}}(1-t)^{\frac{1}{2}}dt}{\int_0^1 t^{\frac{n-1}{2}}(1-t)^{\frac{1}{2}}dt}.\end{aligned}$$

The last term is the CDF of the Beta distribution. Using the fact that

$$\mathbb{P}\left(X_j \in \mathcal{C}(X_{j'},h)\right) \quad = \quad \mathbb{P}\left(X_{j'} \in \mathcal{C}(X_j,h)\right)$$

the union bound, and the independence of the X_j's, the probability that $X_j \in \mathcal{C}(X_{j'},h)$ for some (j,j') in $\{1,\ldots,p_0\}^2$ can be bounded as follows

$$\begin{aligned}\mathbb{P}\left(\cup_{j\neq j'=1}^{p_0}\{X_j \in \mathcal{C}(X_{j'},h)\}\right) &= \mathbb{P}\left(\cup_{j<j'=1}^{p_0}\{X_j \in \mathcal{C}(X_{j'},h)\}\right) \\ &\leq \sum_{j<j'=1}^{p_0}\mathbb{P}\left(\{X_j \in \mathcal{C}(X_{j'},h)\}\right) \\ &= \sum_{j<j'=1}^{p_0}\mathbf{E}\left[\mathbb{P}\left(\{X_j \in \mathcal{C}(X_{j'},h)\} \mid X_{j'}\right)\right] \\ &= \frac{p_0(p_0-1)}{2}\int_0^{2h-h^2} t^{\frac{n-1}{2}}(1-t)^{\frac{1}{2}}dt.\end{aligned}$$

Our next task is to choose h so that

$$\frac{p_0(p_0-1)}{2}\int_0^{2h-h^2} t^{\frac{n-1}{2}}(1-t)^{\frac{1}{2}}dt \quad \leq \quad p_0^{-n}.$$

Let us make the following crude approximation

$$\frac{p_0(p_0-1)}{2}\int_0^{2h-h^2} t^{\frac{n-1}{2}}(1-t)^{\frac{1}{2}}dt \quad \leq \quad \frac{p_0^2}{2}(2h)^{\frac{n-1}{2}}(2h-0).$$

Thus, taking

$$h \quad \geq \quad \frac{1}{2}\,\exp\left(-2\,\left(\log(p_0)+\frac{\log\left(p_0\right)-\log(2))}{n+1}\right)\right)$$

will work. Moreover, since $p_0 \geq 2$, we deduce that

$$\mu(X_{\tilde{I}}) \quad \leq \quad \frac{1}{2}\,p_0^{-2} \tag{3.-12}$$

with probability at least $1-p_0^{-n}$.

Step 2: The norm of $X_{\tilde{I}}$. The norm of any submatrix X_S with n rows and κs columns of X has the following variational representation

$$\|X_S\| \quad = \quad \max_{\substack{v\in\mathbb{R}^n,\ \|v\|=1\\ w\in\mathbb{R}^{\kappa s},\ \|w\|=1}} v^t X_S w.$$

We will use an easy ε-net argument to control this norm. For any $v\in\mathbb{R}^n$, v^tX_j, $j\in S$ is a sub-Gaussian random variable satisfying

$$\mathbb{P}\left(|v^tX_j|\geq u\right) \quad \leq \quad 2\exp\left(-cn\ u^2\right),$$

for some constant c. Therefore, using the fact that $\|w\|=1$, we have that

$$\mathbb{P}\left(\left|\sum_{j\in S} v^tX_S w\right|\geq u\right) \quad \leq \quad 2\exp\left(-cn\ u^2\right).$$

Let us recall two useful results of Rudelson and Vershynin. The first one gives a bound on the covering number of spheres.

Proposition 3.1 *([22, Proposition 2.1]). For any positive integer d, there exists an ε-net of the unit sphere of $\mathbb{R}^d$ of cardinality*

$$2d\left(1+\frac{2}{\varepsilon}\right)^{d-1}.$$

The second controls the approximation of the norm based on an ε-net.

Proposition 3.2 *([22, Proposition 2.2]). Let $\mathcal{N}$ be an ε-net of the unit sphere of $\mathbb{R}^d$ and let $\mathcal{N}'$ be an ε'-net of the unit sphere of $\mathbb{R}^{d'}$. Then for any linear operator $A:\mathbb{R}^d\mapsto\mathbb{R}^{d'}$, we have*

$$\|A\| \quad \leq \quad \frac{1}{(1-\varepsilon)(1-\varepsilon')}\sup_{\substack{v\in\mathcal{N}\\ w\in\mathcal{N}'}}|v^tAw|.$$

Let $\mathcal{N}$ (resp. $\mathcal{N}'$) be an ε-net of the unit sphere of $\mathbb{R}^{\kappa s}$ (resp. of $\mathbb{R}^n$). On the other hand, we have that

$$\begin{aligned}\mathbb{P}\left(\sup_{\substack{v\in\mathcal{N}\\ w\in\mathcal{N}'}}|v^tAw|\geq u\right) &\leq 2|\mathcal{N}||\mathcal{N}'|\exp\left(-cn\ u^2\right),\\ &\leq 8\ n\kappa s\left(1+\frac{2}{\varepsilon}\right)^{n+\kappa s-2}\exp\left(-cn\ u^2\right),\end{aligned}$$

which gives

$$\mathbb{P}\left(\sup_{\substack{v\in\mathcal{N}\\ w\in\mathcal{N}'}}|v^tAw|\geq u\right) \leq 8\ \frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\exp\left(-\left(cn\ u^2-(n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right)\right)\right).$$

Using Proposition (5.2), we obtain that

$$\mathbb{P}\left(\|X_S\|\geq u\right) \leq \mathbb{P}\left(\frac{1}{(1-\varepsilon)^2}\sup_{\substack{v\in\mathcal{N}\\ w\in\mathcal{N}'}}|v^tAw|\geq u\right).$$

Thus, we obtain

$$\mathbb{P}\left(\|X_S\|\geq u\right) \leq 8\ \frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\exp\left(-\left(cn\ (1-\varepsilon)^4\ u^2-(n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right)\right)\right).$$

To conclude, let us note that

$$\begin{aligned}\mathbb{P}\left(\|X_{\tilde{I}}\|\geq u\right) &\leq \mathbb{P}\left(\max_{\substack{S\subset\{1,\dots,p_0\}\\ |S|=\kappa s}}\|X_S\|\geq u\right)\\ &\leq \binom{p_0}{\kappa s}\ 8\ \frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\exp\left(-\left(cn\ (1-\varepsilon)^4\ u^2-(n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right)\right)\right).\end{aligned}$$

and using the fact that

$$\binom{p_0}{\kappa s} \leq \left(\frac{e\ p_0}{\kappa s}\right)^{\kappa s},$$

one finally obtains

$$\mathbb{P}\left(\|X_{\tilde{I}}\|\geq u\right) \leq 8\ \exp\left(-\left(cn\ (1-\varepsilon)^4\ u^2-(n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right)-\kappa s\log\left(\frac{e\ p_0}{\kappa s}\right)-\log\left(\frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\right)\right)\right).$$

The right hand side term will be less than $8p_0^{-n}$ when

$$n\log(p_0) \leq cn\ (1-\varepsilon)^4\ u^2-(n+\kappa s)\log\left(1+\frac{2}{\varepsilon}\right)-\kappa s\log\left(\frac{e\ p_0}{\kappa s}\right)-\log\left(\frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\right).$$

This happens if

$$u^2 \geq \frac{1}{c(1-\varepsilon)^4}\left(n\frac{\log(p_0)}{n}+\left(1+\frac{\kappa s}{n}\right)\log\left(1+\frac{2}{\varepsilon}\right)+\frac{\kappa s}{n}\log\left(\frac{e\ p_0}{\kappa s}\right)+\frac{1}{n}\log\left(\frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\right)\right).$$

Notice that

$$\begin{aligned}\left(1+\frac{\kappa s}{n}\right)\log\left(1+\frac{2}{\varepsilon}\right)+\frac{\kappa s}{n}\log\left(\frac{e}{\kappa s}\right)+\frac{1}{n}\log\left(\frac{n\kappa s\ \varepsilon^2}{(2+\varepsilon)^2}\right) \\ \leq (1+C_\kappa)\log\left(1+\frac{2}{\varepsilon}\right)+C_\kappa+\frac{1}{n}\log\left(\frac{C_\kappa n^2}{4}\right), \\ \leq K_\varepsilon\ \frac{6}{\sqrt{2\pi}},\end{aligned} \tag{3.-27}$$

since $n \geq 1$. Now, since

$$\frac{6}{\sqrt{2\pi}} \quad \leq \log(p_0) \leq \quad \frac{n+\kappa s}{n}\log(p_0),$$

we finally obtain

$$\mathbb{P}\left(\|X_{\tilde{I}}\| \geq \frac{1+K_\varepsilon}{c(1-\varepsilon)^4}\frac{n+\kappa s}{n}\log(p_0)\right) \quad \leq \quad \frac{8}{p_0^n}. \tag{3.-29}$$

Step 3. We will use the following lemma on the distance to identity of randomly selected submatrices.

Lemma 3.3 *Let $r \in (0,1)$. Let n, κ and s satisfy conditions (2.5) and (2.4) assumed in Proposition 2.2. Let $\Sigma \subset \{1,\ldots,\kappa s\}$ be a random support with uniform distribution on index sets with cardinal s. Then, with probability greater than or equal to $1-9\ p_0^{-n}$ on X, the following bound holds:*

$$\mathbb{P}\left(\|X_\Sigma^t X_\Sigma - I_s\| \geq r \mid X\right) \quad < \quad 1. \tag{3.-28}$$

extbfProof. See Appendix. □

Taking $r = 1-\rho_-$, we conclude from Lemma 3.3 that, for any s satisfying (2.6), there exists a subset $\tilde{J}$ of $\tilde{I}$ with cardinal s such that

$$\sigma_{\min}\left(X_{\tilde{J}}\right) \geq \rho_-.$$

The supremum over an ε-net

Recalling Proposition 5.1, there exists an ε-net $\mathcal{N}$ covering the unit sphere in $\mathbb{R}^n$ with cardinal

$$|\mathcal{N}| \quad \leq \quad 2n\left(1+\frac{2}{\varepsilon}\right)^{n-1}.$$

Combining this with (3.-1), we have that

$$\begin{aligned}\mathbb{P}\left(\sup_{v\in\mathcal{N}}\inf_{I\subset\mathcal{S}_{s,\rho_-}}\|X_I^t v\| \geq \frac{8}{e^{2\ln(2)}}\ \frac{\sqrt{\pi}}{}\ \frac{n\ (n-2)^{1/2}}{(n-3)^{3/2}}\ \frac{\log(p_0)}{p_0}\right) \\ \leq 2n\left(1+\tfrac{2}{\varepsilon}\right)^{n-1}\ p_0^{-n}+9\ p^{-n}.\end{aligned} \tag{3.-30}$$

From the ε-net to the whole sphere

For any v', one can find $v \in \mathcal{N}$ with $\|v' - v\|_2 \leq \varepsilon$. Thus, we have

$$\begin{aligned} \|X_I^t v'\|_\infty &\leq \|X_I^t v\|_\infty + \|X_I^t (v' - v)\|_\infty \\ &\leq \|X_I^t v\|_\infty + \max_{j \in I} |\langle X_j, (v' - v) \rangle| \\ &\leq \|X_I^t v\|_\infty + \max_{j \in I} \|X_j\|_2 \|v' - v\|_2 \\ &\leq \|X_I^t v\|_\infty + \varepsilon. \end{aligned} \tag{3.-32}$$

Taking

$$\varepsilon = 80 \, \frac{\log(p_0)}{p_0},$$

we obtain from (3.-32) and (3.-30) that

$$\begin{aligned} &\mathbb{P}\left(\sup_{\|v\|_2=1} \inf_{I \subset \mathcal{S}_{s,\rho_-}} \|X_I^t v\| \geq 80 \, \tfrac{\log(p_0)}{p_0} \right) \\ &\leq 20 \, n \left(1 + \tfrac{p_0}{80 \log(p_0)} \right)^{n-1} \, p_0^{-n} + 9 \, p_0^{-n} \end{aligned}$$

and thus,

$$\begin{aligned} &\mathbb{P}\left(\sup_{\|v\|_2=1} \inf_{I \subset \mathcal{S}_{s,\rho_-}} \|X_I^t v\| \geq 80 \, \tfrac{\log(p_0)}{p_0} \right) \\ &\leq 5 \, \tfrac{n}{p_0 \, \log(p_0)^{n-1}} + 9 \, p_0^{-n}, \end{aligned}$$

for $p_0 \geq \exp(6/\sqrt{2\pi})$.

Proof of Theorem 3.1

Optimality conditions

The optimality conditions for the LASSO are given by

$$-X^t(y - X\hat{\beta}) + \lambda g = 0 \tag{3.-36}$$

for some $g \in \partial \left(\|\cdot\|_1 \right)_{\hat{\beta}}$. Thus, we have

$$X^t X(\hat{\beta} - \beta) = X^t \varepsilon - \lambda g. \tag{3.-35}$$

from which one obtains that, for any index set $H \subset \{1, \ldots, p\}$ with cardinal s,

$$\left\| X_H^t X(\beta - \hat{\beta}) \right\|_\infty \leq \lambda + \left\| X_H^t \varepsilon \right\|_\infty, \tag{3.-34}$$

The support of $\hat{\beta}$

As is well known, even when the solution of the LASSO optimization problem is not unique, there always exists a vector $\hat{\beta}$ whose support has cardinal n.

A bound on $\|X_H^t X_S(\beta_S - \hat{\beta}_S)\|_\infty$

The argument is divided into three steps.

First step. Equation (3.-34) implies that

$$\left\|X_H^t X_S(\beta_S - \hat{\beta}_S)\right\|_\infty \quad \leq \quad \lambda + \left\|X_H^t \varepsilon\right\|_\infty + \left\|X_H^t X_{S^c}(\beta_{S^c} - \hat{\beta}_{S^c})\right\|_\infty . \tag{3.-33}$$

Second step. We now choose H as a solution of the following problem

$$\vartheta \quad = \quad \min_{\substack{I \subset \{1,\ldots,p\} \\ |I|=s}} \max_{j \in I} |\langle X_j, X_{S^c}(\beta_{S^c} - \hat{\beta}_{S^c})\rangle|$$

subject to

$$\sigma_{\min}(X_I) \quad \geq \quad \rho_- .$$

By Definition 2.1,

$$\vartheta \quad \leq \quad \gamma_{s,\rho_-}(X)\ \|X_{S^c}(\beta_{S^c} - \hat{\beta}_{S^c})\|_2$$

and thus,

$$\begin{aligned} \vartheta \quad &\leq \quad \gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c}) \|\beta_{S^c} - \hat{\beta}_{S^c}\|_2 \\ &\leq \quad \gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c}) \|\beta_{S^c} - \hat{\beta}_{S^c}\|_1 \end{aligned}$$

which gives

$$\vartheta \quad \leq \quad \gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c}) \left(\|\beta_{S^c}\|_1 + \|\hat{\beta}_{S^c}\|_1\right). \tag{3.-37}$$

Third step. Combining (3.-33) and (2.3), we obtain

$$\left\|X_H^t X_S(\beta_S - \hat{\beta}_S)\right\|_\infty \quad \leq \quad \lambda + \left\|X_H^t \varepsilon\right\|_\infty + \gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c}) \left(\|\beta_{S^c}\|_1 + \|\hat{\beta}_{S^c}\|_1\right).$$

Using the fact that

$$\left\|X_H^t \varepsilon\right\|_\infty \quad \leq \quad \left\|X^t \varepsilon\right\|_\infty \tag{3.-37}$$

and since

$$\mathbb{P}\left(\left\|X^t \varepsilon\right\|_\infty \geq \sigma\sqrt{2\alpha\ \log(p) + \log(2p)}\right) \quad \leq \quad p^{-\alpha},$$

we obtain that

$$\begin{aligned} \left\|X_H^t X_S(\beta_S - \hat{\beta}_S)\right\|_\infty \quad \leq \quad & \lambda + \sigma\sqrt{(2\alpha+1)\ \log(p) + \log(2)} \\ & + \gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c}) \left(\|\beta_{S^c}\|_1 + \|\hat{\beta}_{S^c}\|_1\right) \end{aligned} \tag{3.-38}$$

with probability greater than $1 - p^\alpha$.

A basic inequality

The definition of $\hat{\beta}$ gives

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \quad \leq \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Therefore, we have that

$$\frac{1}{2}\|\varepsilon - X(\hat{\beta} - \beta)\|_2^2 + \lambda\|\hat{\beta}\|_1 \quad \leq \quad \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\|\beta\|_1$$

which implies that

$$\begin{aligned}\frac{1}{2}\|X(\hat{\beta} - \beta)\|_2^2 \quad &\leq \quad \langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle + \langle \varepsilon, X_{S^c}(\hat{\beta}_{S^c} - \beta_{S^c})\rangle \\ &\quad +\lambda\left(\|\beta_S\|_1 - \|\hat{\beta}_S\|_1\right) - \lambda\|\hat{\beta}_{S^c}\|_1 + \lambda\|\beta_{S^c}\|_1.\end{aligned}$$

This can be further written as

$$\begin{aligned}\frac{1}{2}\|X(\hat{\beta} - \beta)\|_2^2 \quad &\leq \quad \langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle + \langle X_{S^c}^t\varepsilon, \hat{\beta}_{S^c} - \beta_{S^c}\rangle \qquad (3.\text{-}41) \\ &\quad +\lambda\left(\|\beta_S\|_1 - \|\hat{\beta}_S\|_1\right) - \lambda\|\hat{\beta}_{S^c}\|_1 + \lambda\|\beta_{S^c}\|_1. \qquad (3.\text{-}40)\end{aligned}$$

Control of $\langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle$

The argument is divided into two steps.

First step. We have

$$\begin{aligned}\langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle \quad &= \quad \langle X_S^t\varepsilon, \hat{\beta}_S - \beta_S\rangle \\ &\leq \quad \left\|X_S^t\varepsilon\right\|_\infty \|\hat{\beta}_S - \beta_S\|_1 \\ &\leq \quad \sqrt{s}\ \left\|X_S^t\varepsilon\right\|_\infty \|\hat{\beta}_S - \beta_S\|_2\end{aligned}$$

and, using the fact that $\sigma_{\min}(X_H) \geq \rho_-$,

$$\langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle \quad \leq \quad \frac{s}{\rho_-\ \sigma_{\min}(X_S)}\ \left\|X_S^t\varepsilon\right\|_\infty \|X_H^t X_S(\hat{\beta}_S - \beta_S)\|_\infty.$$

Second step. Since the columns of X have unit ℓ_2-norm, we have

$$\mathbb{P}\left(\left\|X_S^t\varepsilon\right\|_\infty \geq \sigma\sqrt{2\alpha\ \log(p) + \log(2s)}\right) \quad \leq \quad p^{-\alpha},$$

which implies that

$$\langle \varepsilon, X_S(\hat{\beta}_S - \beta_S)\rangle \quad \leq \quad \frac{s\ \sigma\sqrt{2\alpha\ \log(p) + \log(2s)}}{\rho_-\ \sigma_{\min}(X_S)}\ \|X_H^t X_S\left(\hat{\beta}_S - \beta_S\right)\|_\infty \qquad (3.\text{-}44)$$

with probability at least $1 - p^{-\alpha}$.

Control of $\langle X^t_{S^c}\varepsilon, \hat{\beta}_{S^c} - \beta_{S^c}\rangle$

We have

$$\langle X^t_{S^c}\varepsilon, \hat{\beta}_{S^c} - \beta_{S^c}\rangle \quad \leq \quad \left\|X^t_{S^c}\varepsilon\right\|_\infty \|\hat{\beta}_{S^c} - \beta_{S^c}\|_1. \tag{3.-43}$$

On the other hand, we have

$$\mathbb{P}\left(\left\|X^t_{S^c}\varepsilon\right\|_\infty \geq \sigma\sqrt{2\alpha\ \log(p) + \log(2(p-s))}\right) \quad \leq \quad p^{-\alpha},$$

which, combined with (3.-43), implies that

$$\langle X^t_{S^c}\varepsilon, \hat{\beta}_{S^c} - \beta_{S^c}\rangle \quad \leq \quad \sigma\sqrt{2\alpha\ \log(p) + \log(2(p-s))}\left(\|\hat{\beta}_{S^c}\|_1 + \|\beta_{S^c}\|_1\right)$$

with probability at least $1 - p^{-\alpha}$.

Control of $\|\beta_S\|_1 - \|\hat{\beta}_S\|_1$

The subgradient inequality gives

$$\|\hat{\beta}_S\|_1 - \|\beta_S\|_1 \quad \geq \quad \langle sign(\beta_S), \hat{\beta}_S - \beta_S\rangle.$$

We deduce that

$$\begin{aligned}\|\beta_S\|_1 - \|\hat{\beta}_S\|_1 \quad &\leq \quad \|-sign(\beta_S)\|_\infty \|\hat{\beta}_S - \beta_S\|_1 \\ &\leq \quad \frac{\sqrt{s}}{\rho_-\ \sigma_{\min}(X_S)}\|X^t_H X_S(\hat{\beta}_S - \beta_S)\|_2\end{aligned}$$

which implies

$$\|\beta_S\|_1 - \|\hat{\beta}_S\|_1 \quad \leq \quad \frac{s}{\rho_-\ \sigma_{\min}(X_S)}\ \|X^t_H X_S(\hat{\beta}_S - \beta_S)\|_\infty. \tag{3.-47}$$

Summing up

Combining (3.-41) with (3.-44), (3.-47) and (3.-38), the union bound gives that, with probability $1 - 3p^{-\alpha}$,

$$\begin{aligned}\frac{1}{2}\|X(\hat{\beta} - \beta)\|_2^2 \quad \leq \quad &\frac{s}{\rho_-\ \sigma_{\min}(X_S)}\left(\sigma\sqrt{2\alpha\ \log(p) + \log(2s)} + \lambda\right)\Bigg(\lambda + \sigma\sqrt{(2\alpha+1)\ \log(p) + \log(2)} \\ &\qquad + \gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c})\left(\|\beta_{S^c}\|_1 + \|\hat{\beta}_{S^c}\|_1\right)\Bigg) \\ &+ \sigma\sqrt{2\alpha\ \log(p) + \log(2(p-s))}\left(\|\beta_{S^c}\|_1 + \|\hat{\beta}_{S^c}\|_1\right) \\ &+ \lambda\left(\|\beta_{S^c}\|_1 - \|\hat{\beta}_{S^c}\|_1\right)\end{aligned}$$

which gives,

$$\begin{aligned}\frac{1}{2}\|X(\hat{\beta}-\beta)\|_2^2 \quad \leq \quad & s\,\frac{\lambda+\sigma\sqrt{(2\alpha+1)\ \log(p)+\log(2)}}{\rho_-\ \sigma_{\min}(X_S)}\left(\sigma\sqrt{2\alpha\ \log(p)+\log(2s)}+\lambda\right)\\ &+\left(\frac{s}{\rho_-\ \sigma_{\min}(X_S)}\left(\sigma\sqrt{2\alpha\ \log(p)+\log(2s)}+\lambda\right)\gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c})\right.\\ &\qquad\left.+\sigma\sqrt{2\alpha\ \log(p)+\log(2(p-s))}-\lambda\right)\|\hat{\beta}_{S^c}\|_1\\ &+\left(\frac{s}{\rho_-\ \sigma_{\min}(X_S)}\left(\sigma\sqrt{2\alpha\ \log(p)+\log(2s)}+\lambda\right)\gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c})\right.\\ &\qquad\left.+\sigma\sqrt{2\alpha\ \log(p)+\log(2(p-s))}+\lambda\right)\|\beta_{S^c}\|_1.\end{aligned}$$

Using the assumption that $s \leq \nu n$, we obtain

$$\begin{aligned}\frac{1}{2}\|X(\hat{\beta}-\beta)\|_2^2 \quad \leq \quad & s\,\frac{\lambda+\sigma\sqrt{(2\alpha+1)\ \log(p)+\log(2)}}{\rho_-\ \sigma_{\min}(X_S)}\left(\sigma\sqrt{2\alpha\ \log(p)+\log(2\nu n)}+\lambda\right)\\ &+\left(\frac{\nu n}{\rho_-\ \sigma_{\min}(X_S)}\left(\sigma\sqrt{2\alpha\ \log(p)+\log(2\nu n)}+\lambda\right)\gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c})\right.\\ &\qquad\left.+\sigma\sqrt{(2\alpha+1)\ \log(p)+\log(2)}-\lambda\right)\|\hat{\beta}_{S^c}\|_1\\ &+\left(\frac{\nu n}{\rho_-\ \sigma_{\min}(X_S)}\left(\sigma\sqrt{2\alpha\ \log(p)+\log(2\nu n)}+\lambda\right)\gamma_{s,\rho_-}(X)\ \sigma_{\max}(X_{S^c})\right.\\ &\qquad\left.+\sigma\sqrt{2\alpha\ \log(p)+\log(2(p))}+\lambda\right)\|\beta_{S^c}\|_1.\end{aligned}$$

Since, as recalled in Section 3, the support of $\hat{\beta}$ has cardinal less than or equal to n, we have

$$\sigma_{\max}(X_{S^c}) \quad \leq \quad \max_{\substack{T\subset\{1,\ldots,p\}\\ |T|\leq n}} \sigma_{\max}(X_T),$$

and the proof is completed.

4 A simple trick when γ_{s,ρ_-} is unknown: appending a random matrix

We have computed the index γ_{s,ρ_-} for the random matrix with independent columns uniformly distributed on the unit sphere of $\mathbb{R}^n$ in Theorem 2.2. The goal of this section is to show that this result can be used in a simple trick in order to obtain prediction bounds similar to [14, Theorem 2.1] without conditions on the design matrix X.

This idea is of course to use Theorem 3.1 above. However, the values of $\sigma_{\min}(X_S)$ and $\sigma_{\max}(X_{S^c})$ are of course usually not known ahead of time and we have to provide easy to compute bounds for these quantities. The coherence $\mu(X)$ can be used for this purpose.

Indeed, for any positive integer $t \leq p$ and any $T \subset \{1, \ldots, p\}$ with $|T| = t$, we have

$$\begin{aligned} \mu(X) &= \|X^t X - I\|_{1,1} \\ &= \max_{\|w\|_\infty = 1} \max_{\|w'\|_1 = 1} w^t (X^t X - I) w' \\ &\geq \frac{1}{\sqrt{t}} \max_{\substack{\|w\|_2 = 1 \\ \|w\|_0 = t}} \max_{\substack{\|w'\|_2 = 1 \\ \|w'\|_0 = t}} w^t (X^t X - I) w'. \end{aligned}$$

Thus, we obtain that

$$1 - \mu(X)\sqrt{t} \quad \leq \sigma_{\min}(X_T) \leq \sigma_{\max}(X_T) \leq \quad 1 + \mu(X)\sqrt{t}.$$

However, the lower bound on $\sigma_{\min}(X_S)$ obtained in this manner may not be accurate enough. More precise, polynomial time computable, bounds have been devised in the litterature. The interested reader can find a very useful Semidefinite relaxation of the problem of finding the worst possible value of $\sigma_{\min}(X_T)$ over all subsets T of $\{1, \ldots, p\}$ with a given cardinal (related to the Restricted Isometry Constant) in [10].

Assuming we have a polynomial time computable a priori bound $\sigma^*_{\min}$ on $\sigma_{\min}(X_T)$ (resp. $\sigma^*_{\max}$ on $\max_{\substack{T \subset \{1,\ldots,p\} \\ |T| \leq n}} \sigma_{\max}(X_T)$), our main result for the case of general design matrices is the following theorem.

Theorem 4.1 *Let X be an matrix in $\mathbb{R}^{n \times p}$ with ℓ_2-normalized columns and let X_0 be a random matrix with independent columns uniformly distributed on the unit sphere of $\mathbb{R}^n$. Let $X_\sharp$ denote the matrix corresponding to the concatenation of X and X_0, i.e. $X_\sharp = [X, X_0]$. Let $\hat{\beta}_\sharp$ denote the LASSO estimator with X replaced with $X_\sharp$ in (1.0). Let $\rho_- \in (0,1)$. Let ν be a positive real. Assume that p_0 is such that*

$$80 \, \frac{\log(p_0)}{p_0} \quad < \quad L \, \rho_- \frac{\sigma^*_{\min}}{\nu n \, \sigma^*_{\max}} \tag{4.-65}$$

for some $L \in (0,1)$. Assume moreover that p_0 is sufficiently large so that the second inequality in (2.5) is satisfied. Assume that β has support S with cardinal s and that

$$\lambda \quad \geq \quad \sigma \left(B'_{X,\nu,\rho_-} \, \sigma^*_{\max} \sqrt{2\alpha \, \log(p + p_0) + \log(2\nu n)} + \sqrt{(2\alpha + 1) \, \log(p + p_0) + \log(2)} \right)$$

with

$$B'_{X,\nu,\rho_-} \quad = \quad \frac{\nu n \, \gamma_{\nu n, \rho_-}(X)}{\rho_- \, \sigma^*_{\min} - \nu n \, \gamma_{\nu n, \rho_-}(X) \, \sigma^*_{\max}}. \tag{4.-65}$$

Assume that s satisfies the first inequality in (2.5) and that $s \leq \nu n$. Then, with probability greater than $1 - p^{-\alpha} - 9 p_0^{-n} - 20 \, \frac{n}{\log(p_0)^{n-1}} p_0^{-1}$, we have

$$\frac{1}{2} \|X(\hat{\beta}_\# - \beta)\|_2^2 \quad \leq \quad s \, C'_{n,p,\rho_-,\alpha,\nu,\lambda} \tag{4.-64}$$

with

$$C'_{n,p,\rho_-,\alpha,\nu,\lambda} \quad = \quad \frac{\lambda + \sigma\sqrt{(2\alpha + 1) \, \log(p + p_0) + \log(2)}}{\rho_- \, \sigma^*_{\min}} \left(\sigma \sqrt{2\alpha \, \log(p + p_0) + \log(2\nu n)} + \lambda \right)$$

extbfProof. Since the index $\gamma_{s,\rho-}$ does not increase after appending a matrix with ℓ_2-normalized columns, the matrix $X_\#$ has at most the same index as that of X_0. Then (4.-65) ensures that the index $\gamma_{s,\rho-}(X_\#)$ is sufficiently small. The rest of the proof is identical to the proof of Theorem 3.1. □

5 Proof of Lemma 3.3

For any index set $S \subset \{1, \dots, \kappa s\}$ with cardinal s, define R_S as the diagonal matrix with

$$(R_S)_{i,i} \quad = \quad \begin{cases} 1 \text{ if } i \in S, \\ 0 \text{ otherwise.} \end{cases}$$

Notice that we have

$$\left\| X_S^t X_S - I \right\| \quad = \quad \|R_S H R_S\|$$

with $H = X^t X - I$. In what follows, R_δ simply denotes a diagonal matrix with i.i.d. diagonal components δ_j, $j = 1, \dots, \kappa s$ with Bernoulli $B(1, 1/\kappa)$ distribution. Let R' be an independent copy of R. Assume that S is drawn uniformly at random among index sets of $\{1, \dots, \kappa s\}$ with cardinal s. By an easy Poissonization argument, similar to [14, Claim (3.29) p.2173], we have that

$$\mathbb{P}\left(\|R_s H R_s\| \geq r\right) \ \leq \ 2\, \mathbb{P}\left(\|RHR\| \geq r\right), \tag{5.-67}$$

and by Proposition 4.1 in [13], we have that

$$\mathbb{P}\left(\|RHR\| \geq r\right) \quad \leq \quad 36\, \mathbb{P}\left(\|RHR'\| \geq r/2\right). \tag{5.-66}$$

In order to bound the right hand side term, we will use [13, Proposition 4.2]. Set $r' = r/2$. Assuming that $\kappa \frac{r'^2}{e} \geq u^2 \geq \frac{1}{\kappa}\|X\|^4$ and $v^2 \geq \frac{1}{\kappa}\|X\|^2$, the right hand side term can be bounded from above as follows:

$$\mathbb{P}\left(\|RHR'\| \geq r'\right) \quad \leq \quad 3\ \kappa s\ \mathcal{V}(s, [r', u, v]), \tag{5.-65}$$

with

$$\mathcal{V}(s, [r', u, v]) \quad = \quad \left(e \frac{1}{\kappa} \frac{u^2}{r'^2}\right)^{\frac{r'^2}{v^2}} + \left(e \frac{1}{\kappa} \frac{\|M\|^4}{u^2}\right)^{u^2/\|M\|^2} + \left(e \frac{1}{\kappa} \frac{\|M\|^2}{v^2}\right)^{v^2/\mu(M)^2}.$$

Using (3.-12) and (3.-29), we deduce that with probability at least $1 - 8p_0^{-n} - p_0^{-n}$, we have

$$\begin{aligned} \mathcal{V}(s, [r', u, v]) \quad = \quad & \left(e \frac{1}{\kappa} \frac{u^2}{r'^2}\right)^{\frac{r'^2}{v^2}} + \left(e \frac{1}{\kappa} \frac{\left(\frac{1+K_\varepsilon}{c(1-\varepsilon)^4} \frac{n+\kappa s}{n} \log(p_0)\right)^4}{u^2} \right)^{\frac{u^2}{\left(\frac{1+K_\varepsilon}{c(1-\varepsilon)^4} \frac{n+\kappa s}{n} \log(p_0)\right)^2}} \\ & + \left(e \frac{1}{\kappa} \frac{\left(\frac{1+K_\varepsilon}{c(1-\varepsilon)^4} \frac{n+\kappa s}{n} \log(p_0)\right)^2}{v^2} \right)^{\frac{v^2}{\frac{1}{2} p_0^{-2}}}. \end{aligned}$$

Take κ, u and v such that

$$\begin{aligned} v^2 \quad &= \quad r'^2 \ \frac{1}{\log(C_\kappa\ n)} \\ u^2 \quad &= \quad C_{\mathcal{V}} \left(\frac{1 + K_\varepsilon}{c(1-\varepsilon)^4} \frac{n + \kappa s}{n} \log(p_0)\right)^2, \\ \kappa \quad &\geq \quad e^3 \frac{C_{\mathcal{V}}}{r'^2} \left(\frac{1 + K_\varepsilon}{c(1-\varepsilon)^4} \frac{n + \kappa s}{n} \log(p_0)\right)^2 \end{aligned}$$

for some $C_{\mathcal{V}}$ possibly depending on s. Since $\kappa s \leq C_\kappa n$, this implies in particular that

$$\kappa \quad \geq \quad e^3 \, \frac{C_{\mathcal{V}}}{r'^2} \, \left(\frac{(1+K_\varepsilon)(1+C_\kappa)}{c(1-\varepsilon)^4} \log(p_0) \right)^2 . \tag{5.-73}$$

Thus, we obtain that

$$\mathcal{V}(s,[r',u,v]) \quad = \quad \left(\frac{1}{e^2}\right)^{\log(C_\kappa n)} + \left(\frac{r'^2}{e^2 \; C_{\mathcal{V}}^2}\right)^{C_{\mathcal{V}}} + \left(\frac{\log(C_\kappa n)}{e^2 \; C_{\mathcal{V}}}\right)^{\frac{2r'^2 \; p_0^2}{\log(C_\kappa n)}} .$$

Using (5), (5.-66) and (5.-65), we obtain that

$$\mathbb{P}\left(\|R_s H R_s\| \geq r'\right) \quad \leq \quad 2 \times 36 \times 3 \times \kappa s \left(\left(\frac{1}{e^2}\right)^{\log(C_\kappa n)} + \left(\frac{r'^2}{e^2 \; C_{\mathcal{V}}^2}\right)^{C_{\mathcal{V}}} + \left(\frac{\log(C_\kappa n)}{e^2 \; C_{\mathcal{V}}}\right)^{\frac{2r'^2 \; p_0^2}{\log(C_\kappa n)}} \right) .$$

Take

$$C_{\mathcal{V}} \quad = \quad \log(C_\kappa n) \tag{5.-74}$$

and, since $p_0 > 1$ and $r \in (0,1)$, we obtain

$$\begin{aligned} &\mathbb{P}\left(\|R_s H R_s\| \geq r'\right) \\ &\qquad \leq 2 \times 36 \times 3 \times \kappa s \left(\left(\frac{1}{e^2}\right)^{\log(C_\kappa n)} + \left(\frac{r'^2}{e^2 \; \log^2(C_\kappa n)}\right)^{\log(C_\kappa n)} + \left(\frac{1}{e^2}\right)^{\frac{2r'^2 \; p_0^2}{\log(C_\kappa n)}} \right) \end{aligned} \tag{5.-74}$$

Replace r' by $r/2$. Since it is assumed that $n \geq \exp(r/2)/C_\kappa$ and $p_0 \geq \sqrt{2}\log(C_\kappa n)/r$, it is sufficient to impose that

$$C_\kappa^2 n^2 \quad \geq \quad (2 \times 36 \times 3 \times \kappa s \times 3)^{\frac{1}{\log(e^2)}} ,$$

in order for the right hand side of (5.-74) to be less than one. Since $\kappa s \leq C_\kappa n$, it is sufficient to impose that

$$C_\kappa^2 n^2 \quad \geq \quad 2 \times 36 \times 3 \times \; C_\kappa n \times 3,$$

or equivalently,

$$C_\kappa n \quad \geq \quad 2 \times 36 \times 3 \times 3.$$

This is implied by (2.5) in the assumptions. On the other hand, combining (5.-73) and (5.-74) implies that one can take

$$\kappa \quad = \quad \frac{4e^3}{r^2} \, \left(\frac{(1+K_\varepsilon)(1+C_\kappa)}{c(1-\varepsilon)^4} \right)^2 \log^2(p_0) \log(C_\kappa n),$$

which is nothing but (2.4) in the assumptions.

Bibliography

[1] Akaike, H., A new look at the statistical model identification, IEEE Trans. Automat. Control, (1974) 19, 716–723. (p. 229).

[2] Bickel, P. J., Ritov, Y., Tsybakov, A. B., Simultaneous analysis of lasso and Dantzig selector. Ann. Statist. 37 (2009), no. 4, 1705–1732. (p. 229).

[3] Bunea, F., Tsybakov, A. and Wegkamp, M., Sparsity oracle inequalities for the LASSO, 2007, The Electronic Journal of Statistics, 169 – 194. (p. 229).

[4] Bühlmann, P., van de Geer, S., Statistics for High-Dimensional Data, Methods, Theory and Applications, Series: Springer Series in Statistics (2011). (p. 229).

[5] van de Geer, S. and Bühlmann, P., On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics (2009) 3, 1360–1392. (p. 229).

[6] Candès, E. J. and Plan, Y. Near-ideal model selection by ℓ_1 minimization. Ann. Statist. 37 (2009), no. 5A, 2145–2177. (pp. 177, 178, 180, 181, 182, 183, 184, 185, 187, 188, 189, 192, 193, 205, 206, 207, 217, 218, 219, 220, 228, 229, 230, 233, 243, 245, 250, 254, 255, 256, 286, 300 et 301).

[7] Chrétien, S. and Darses, S., Invertibility of random submatrices via tail decoupling and a Matrix Chernoff Inequality, Stat. and Prob. Lett., (2012), 82, no. 7, 1479–1487. (pp. 229, 245 et 282).

[8] Chrétien, S. and Darses, S., Sparse recovery with unknown variance: a LASSO-type approach, IEEE Trans. Info. Th., to appear.

[9] Dalalyan, A. and Tsybakov, A., Sparse Regression Learning by Aggregation and Langevin Monte-Carlo, J. Comput. System Sci. 78 (2012), pp. 1423-1443. (p. 229).

[10] A. d'Aspremont, F. Bach and L. El Ghaoui, Optimal Solutions for Sparse Principal Component Analysis. Journal of Machine Learning Research, 9 (2008), pp. 1269-1294. (p. 244).

[11] Dubhashi, D. P. and Panconesi, A., Concentration of measure for the analysis of randomized algorithms. Cambridge University Press, Cambridge, 2009. (p. 233).

[12] The Risk Inflation Criterion for Multiple Regression. Dean P. Foster and Edward I. George. Annals of Statistics, Volume 22, Issue 4 (1994), 1947–1975. (p. 229).

[13] Muirhead, R., Aspects of multivariate statistical theory. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1982. xix+673 pp. (p. 233).

[14] Qi, F, Bounds for the ratio of two gamma functions. J. Inequal. Appl. 2010. (p. 234).

[15] Rigollet, P. and Tsybakov, A., Exponential Screening and optimal rates of sparse estimation, Ann. Statist., 39(2), 731-771. (p. 229).

[16] Rudelson, M. and Vershynin, R., Smallest singular value of a random rectangular matrix. Comm. Pure Appl. Math. 62 (2009), no. 12, 1707–1739. (pp. 236 et 280).

[17] Estimating the dimension of a model, The Ann. of Stat., (1978) 6, 461–464. (p. 229).

[18] Tibshirani, R. Regression shrinkage and selection via the LASSO, J.R.S.S. Ser. B, 58, no. 1 (1996), 267–288. (pp. 177, 228 et 250).

[19] Tropp, J., User friendly tail bounds for sums of random matrices, Foundations of Computational Mathematics, (2012), 12, no.4, 389–434. (p. 229).

[20] van de Geer, S. and Bühlmann, P., On the conditions used to prove oracle results for the Lasso. Electron. J. Stat. 3 (2009) 1360–1392. (pp. 178 et 229).

[21] Wainwright, Martin J., Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory 55 (2009), no. 5, 2183–2202. (pp. 177, 186, 187, 229 et 250).

[22] Zhao, P. and Yu, B., On model selection consistency of Lasso. J. Mach. Learn. Res. 7 (2006), 2541–2563. (pp. 177 et 229).

Chapter K

Mixture model for designs in high dimensional regression and the LASSO

Abstract

The LASSO is a recent technique for variable selection in the regression model

$$y \quad = \quad X\beta + \varepsilon,$$

where $X \in \mathbb{R}^{n\times p}$ and ε is a centered gaussian i.i.d. noise vector $\mathcal{N}(0, \sigma^2 I)$. The LASSO has been proved to perform exact support recovery for regression vectors when the design matrix satisfies certain algebraic conditions and β is sufficiently sparse. Estimation of the vector $X\beta$ has also extensively been studied for the purpose of prediction under the same algebraic conditions on X and under sufficient sparsity of β. Among many other, the coherence is an index which can be used to study these nice properties of the LASSO. More precisely, a small coherence implies that most sparse vectors, with less nonzero components than the order $n/\log(p)$, can be recovered with high probability if its nonzero components are larger than the order $\sigma\sqrt{\log(p)}$. However, many matrices occuring in practice do not have a small coherence and thus, most results which have appeared in the litterature cannot be applied. The goal of this paper is to study a model for which precise results can be obtained. In the proposed model, the columns of the design matrix are drawn from a Gaussian mixture model and the coherence condition is imposed on the much smaller matrix whose columns are the mixture's centers, instead of on X itself. Our main theorem states that $X\beta$ is as well estimated as in the case of small coherence up to a correction parametrized by the maximal variance in the mixture model.

1 Introduction

The goal of the present paper is the study of the high dimensional regression problem $y = X\beta + z$, where $X \in \mathbb{R}^{n\times p}$, with $p \gg n$ and $z \sim \mathcal{N}(0, \sigma^2 I_n)$. For simplicity, we will assume throughout this paper that the columns of X have unit l_2-norm. This problem has been the subject of a great research activity. This high dimensional setting, where more variables are involved than observations, occurs in many different applications such as image processing and denoising, gene expression analysis, and, after slight modifications, time series (filtering) [17], [20], machine learning and especially graphical models [19] and more recently, biochemistry [1]. One of the most popular approaches is the Least Angle

Shrinkage and Selection Operator (LASSO) introduced in [23] for the purpose of variable selection. The LASSO estimator is given as a solution, for $\lambda > 0$, of

$$\hat{\beta} = argmin_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \lambda \|b\|_1. \tag{1.0}$$

Conditions for uniqueness of the minimizer in this last expression are discussed in [16], [21] and [15]. Several other estimators have also been proposed, such as the Dantzig Selector [11] [3] or Message Passing Algorithms [14]. In the sequel, we will focus on the LASSO due to its wide use in various applications.

One of the most surprising and important discoveries from these recent extensive efforts is that, under appropriate assumptions on the design matrix X, and for most regression vectors β, the support of β can be recovered exactly when its size is of the order $n/\log(p)$; see [3], [5], [14], [28] for instance. Moreover, under similar assumptions, the prediction error can be controlled adaptively as a function of the sparsity of β and the noise variance; see for instance [14]. Similar rates can be achieved by other method, involving for instance penalization, but the main advantage of the LASSO over most competitors is that a solution can be obtained in polynomial time, following the definition of complexity theory. A very efficient algorithm is, e.g., [2]. Many implementations are available on the web.

The two main assumptions for achieving these remarkable results are unavoidably imposed on the design matrix X and on the regression vector.

- The regression vector β is assumed to be s-sparse, with support denoted by T, meaning that no more than s of its components are non zero. This can be relaxed to β assumed only compressible, that is approximable by a sparse vector.

- The design matrix is assumed to satify one of many proposed algebraic conditions in the litterature, implying that all singular values of X_S are close to one for any or most given $S \subset \{1, \ldots, n\}$ with $|S| = s'$ for some appropriate choice of s' (often equal to s of $2s$).

Concerning the second point, two main assumptions have been proposed in the litterature. The first is the Restricted Isometry Property [12] [8], which requires that

$$(1-\delta)\|\beta_S\|_2^2 \leq \quad \|X_S\beta_S\|_2^2 \quad \leq (1+\delta)\|\beta_S\|_2^2, \tag{1.1}$$

for $S \subset \{1, \ldots, n\}$ with $|S| = s'$ and all $\beta \in \mathbb{R}^p$. This property is satisfied by high probability for most random matrices with i.i.d. entrees with variance $1/n$ such as Gaussian or Rademacher variables and for $s' \leq C_{rip}\ n/\log(p)$, where the constant C_{rip} depends on the distribution of the individual entrees. Notice that the $1/n$ assumption on the variance and standard concentration bounds imply that the resulting random matrix has almost normalized columns and the normalized avatar will satisfy RIP with unessential modifications of the constants. The RIP has been extensively used in signal processing after the emergence of the so-called Compressed Sensing paradigm [7].

The second assumption which is often considered is the Incoherence Condition, which requires that

$$\mu(X) \quad = \quad \max_{j \neq j'=1}^{p} |\langle X_j, X_{j'} \rangle|$$

is small, e.g. $\mu(X) \leq C_\mu/\log(p)$ as in [14], which is garanteed for random matrices with i.i.d. gaussian entrees with variance $1/n$ in the range $n \geq C_{ic}\log(p)^3$.

The main advantage of the Incoherence Condition over the Restricted Isometry Property is that it can be checked quicky (in $p(p-1)/2$ operations), whereas no-one knows how to check the RIP without enumerating all possible supports $S \subset \{1,\dots,n\}$ with cardinal s'. Such an enumeration would of course take an exponential amount of time to establish. The main relationship between IC and RIP is that it can be proved that under IC, (1.1) holds, not for all, but for most supports $S \subset \{1,\dots,n\}$ with cardinal s', where $s' \leq C_s\, p/(\|X\| \log(p))$, for some constant C_s controlling the proportion of such supports.

The objective of the present paper is to extend the analysis based on the Incoherence Condition to more general situations where X may have a lot of very colinear columns. The main idea is to assume that the columns are drawn from a mixture model of K clusters, and that the set of cluster's centers form a matrix which satisfies the Incoherence Condition.

2 Main results

The mixture model

In order to relax the Incoherence Condition, one needs a model for the design matrix X allowing for a certain amount of correlations between columns while keeping some of the algebraic structure in the same spirit as (1.1) for at least most supports indexing a subset of really pertinent covariates. In what follows, we study such a model, where the columns can be considered as belonging to a family of clusters and the cluster's centers or (an empirical surrogate) is defined to be the pertinent variable. This model is of great interest when many columns are very colinear. In practice, one often observes that the columns of X can be grouped into different clusters such that the dot product of X_j and $X_{j'}$ $j \neq j'$ is close to one if they belong to the same cluster, and very close to zero otherwise. Notice that applying the LASSO for such designs will eventually result into grossly incorrect variable selection. On the other hand confusing a variable for another very correlated variable might not be a real issue as far as prediction is concerned if the clusters are well separated.

Detailed presentation

Let K be the number of clusters in the covariates. Consider a matrix $\mathfrak{C}$ in $\mathbb{R}^{n\times K}$, with small coherence. The columns of the matrix $\mathfrak{C}$ will be the "centers" of each cluster, $k = 1,\dots,K$.

The design matrix will be assumed to derive from a matrix X_o whose columns are drawn from the following procedure. Let $\mathcal{K}$ be randomly drawn among all index subsets of $\{1,\dots,K\}$ with cardinal s^* with uniform distribution. We then assume that, conditionally on $\mathcal{K}$ each column of X_o is drawn from a mixture Φ of K n-dimensional Gaussian distributions, i.e.

$$\Phi(x) \quad = \quad \sum_{k\in\mathcal{K}} \pi_k \phi_k(x),$$

where

$$\phi_k(x) \quad = \quad \frac{1}{(2\pi\mathfrak{s}^2)^{\frac{n}{2}}} \exp\left(-\frac{\|x-\mathfrak{C}_k\|_2^2}{2\mathfrak{s}^2}\right),$$

and $\pi_k \geq 0$, $k \in \mathcal{K}$ and $\sum_{k\in\mathcal{K}} \pi_k = 1$. We will denote by n_k the random number of columns in X_o that were drawn from $\mathcal{N}(\mathfrak{C}_k, \mathfrak{s}^2 I)$, $k = 1,\dots,K_o$. Thus, $\sum_{k\in\mathcal{K}} n_k = p$.

Finally, the matrix X is obtained by column-wise normalization of X_o, i.e. $X_j = X_{o,j}/\|X_{o,j}\|_2$.

Notice that the model could easily be modified in order to more general distributions for $\mathcal{K}$ than the uniform distribution on subsets of $\{1,\dots,K_o\}$ with cardinal s^*.

More notations

For each $j \in \{1,\dots,p\}$, denote by k_j the index of the Gaussian component from which columns j was drawn, and let J_k denote the set of indices of columns drawn from the k^{th} Gaussian component. For any index set $S \in \{1,\dots,p\}$, let $\mathcal{K}_S$ denote the list (with possible repetitions)

$$\mathcal{K}_S \quad = \quad \{k_j \mid j \in S\}\,.$$

The deviation of columns $X_{o,j}$ from center $\mathfrak{C}_{k_j}$ will be denoted by

$$\varepsilon_j \quad = \quad X_{o,j} - \mathfrak{C}_{k_j} \sim \mathcal{N}\left(\mathfrak{C}_{k_j}, \mathfrak{s}^2\right).$$

and the matrix E is defined as

$$E \quad = \quad (\varepsilon_{i,j})_{i\in\{1,\dots,n\},\ j\in\{1,\dots,p\}}.$$

A simple proxy for β

For each $k \in \{1,\dots,K\}$, let j_k^* be the best approximation of the center $\mathfrak{C}_k$ from the set of colums X_j, $j \in J_k$, i.e.

$$j_k^* \quad = \quad argmin_{j\in J_k}\|X_j - \mathfrak{C}_k\|_2.$$

Moreover, set

$$T^* \quad = \quad \{j_k^* \mid k \in \mathcal{K}\}\,.$$

Of course, we have $s^* = |T^*|$.

The vector β^* is defined by

$$\mathfrak{C}_{\mathcal{K}_{T^*}}\beta^*_{T^*} \quad = \quad \mathfrak{C}_{\mathcal{K}_T}\beta_T. \tag{2.-6}$$

A simple expression of β^* can be obtained by taking

$$\beta^*_{j^*} \quad = \quad \sum_{j\in J_{k_{j^*}}\cap T} \beta_j \tag{2.-5}$$

for all $j^* \in T^*$. Moreover, this expression is unique whenever X_{T^*} has rank equal to s^*. In Section 3, we will show that X_{T^*} is indeed non-singular with high probability under appropriate assumptions on T.

Main result

Further notations

In the sequel r will denote a constant in $(1, 1/4)$. The constants ϑ_* et ν will be specified in Assumptions 2.3 below. The constants C_μ, C_{spar} et C_{col} will be used in the Assumptions below:

$$C_\mu \quad = \quad r/(1+\alpha),$$

$$C_{spar} = r^2/((1+\alpha)e^2),$$

$$C_{col} = \frac{1}{2}\left(\frac{\sqrt{2}}{\sqrt{(1-r)(1+\alpha)}} - (1+r)\right).$$

Let C_χ denote a positive constant such that

$$\mathbb{P}\left(\frac{\|G\|_2^2}{\mathfrak{s}^2} \leq u^2\right) \leq C_\chi \left(\frac{u^2}{n}\right)^n$$

where G is a n-dimensional centered and unit-variance i.i.d. gaussian vector. Let us further define

$$r_{\max} = 1 + \mathfrak{s}\left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right), \tag{2.-8}$$

$$\mu_{\max} = \frac{1}{2}\,\mathfrak{s}\left(\sqrt{n} + \sqrt{s} + \sqrt{2\alpha\log(p)}\right), \tag{2.-7}$$

$$\sigma^2_{\max} = \frac{1}{2}\sqrt{s}\;\mathfrak{s}^2, \tag{2.-6}$$

$$r^*_{\max} = \frac{1}{1 - \mathfrak{s}\sqrt{n\;\left(\frac{\alpha\;(1-e^{-1})}{\vartheta_*\;C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}},$$

$$K^2_{n,s^*} = \alpha\;n\;\log(p)\;\left(\frac{\alpha\;(1-e^{-1})}{\vartheta_*\;C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}, \tag{2.-6}$$

$$\mu^*_{\max} = \mathfrak{s}K_{n,s^*},$$

$$\sigma^{*\;2}_{\max} = \frac{\left(\frac{\alpha\;(1-e^{-1})}{\vartheta_*\;C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}{1 - \mathfrak{s}\sqrt{n\;\left(\frac{\alpha\;(1-e^{-1})}{\vartheta_*\;C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}\sqrt{s^*}\;\mathfrak{s}^2, \tag{2.-6}$$

$$C_\int = \int_0^3 \sqrt{\log\left(\frac{3}{\varepsilon'}\right)}\;d\varepsilon'.$$

and

$$C^*_\int = \int_0^3 \sqrt{\log\left(\frac{3}{\varepsilon'}\right)}\;d\varepsilon'.$$

Assumptions

We will make the following assumptions.

Assumptions 2.1

$$p \geq \max\left\{K_o, e^{e^{2-\log(\alpha)}}\right\}.$$

and

$$\log(p) \geq \max\left\{\frac{0.2\cdot r\left(1+1.1\cdot r+0.11\cdot r^2\right)}{0.1\cdot\left(1.1\cdot r+0.11\cdot r^2\right)}, \frac{1.1\cdot r\ \ (1.1+0.11\cdot r)}{\alpha}\right\}.$$

Assumptions 2.2 *Assume that $\mathfrak{C}$ has coherence $\mu(\mathfrak{C})$ satisfying*

$$\mu(\mathfrak{C}) \leq \frac{C_\mu}{\log(p)}. \tag{2.-9}$$

Assumptions 2.3 *There exists a positive real constant ϑ^* and a positive integer ν such that*

$$\min_{j^*\in T^*} |\mathcal{J}_{k_{j^*}}| \geq \vartheta_* \log(p)^\nu.$$

Assumptions 2.4

$$s^* \leq \frac{K_o}{\log p}\ \frac{C_{spar}}{\|\mathfrak{C}\|^2}.$$

Assumptions 2.5

$$n \geq \frac{\alpha+1}{c}\log(p). \tag{2.-10}$$

Remark 2.1 *Assumption 2.5 is to be interpreted with care since the order of magnitude of n is primarily governed by Assumption 2.2 on the coherence of $\mathfrak{C}$. For instance, if $\mathfrak{C}$ comes from a Gaussian i.i.d. random matrix, the coherence will be of the order $\sqrt{\log(K_o)/n}$ as discussed in [14, Section 1.1] and n should be at least of the order $\log(p)^2\log(K_o)$. Notice that this is still less than if X itself had to satisfy the coherence bound, with would implie that n be of the order $\log(p)^3$.*

Assumptions 2.6

$$C_{col} \geq e^2(\alpha+1)\max\{\sqrt{C_{spar}}, C_\mu\}.$$

and

$$\left(C_{col}+(1+1.1\cdot r)\ C_{\mathfrak{s},n,p}\right) \leq \frac{1}{2}\sqrt{\frac{\log(p)\ (1-r^*)^2}{(\alpha\log(p)-\log(2))\,2}}.$$

Assumptions 2.7 $\mathfrak{s}$ $\left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right) \leq 1/2$

$$\mathfrak{s} \leq C_{\mathfrak{s},n,p}\frac{1}{\sqrt{\log(p)}\ \left(\sqrt{n}+\sqrt{\frac{\alpha+1}{c}\log(p)}\right)}. \tag{2.-11}$$

for any $C_{\mathfrak{s},n,p}$ such that and

$$C_{\mathfrak{s},n,p} \leq \min\left\{0.1 \cdot \frac{r}{\sqrt{\alpha\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}; \frac{1}{2}\sqrt{\log(p)}\right\}.$$

Assumptions 2.8

$$\|\beta_T\|_2^2 \geq \frac{2\ \alpha \log(p) n\ \sigma_{\max}^2}{\frac{4}{9}\frac{\alpha^2}{}\ \mu_{\max}^2 \log^2(p) - 12\ C_{\int}\ \mu_{\max}\ r_{\max}\ \mathfrak{s}\sqrt{n}}.$$

Assumptions 2.9

$$\|{\beta^*}_{T^*}\|_2^2 \geq \frac{2\alpha \log(p) n\ {\sigma^*_{\max}}^2}{\frac{4}{9}\frac{\alpha^2}{}\ {\mu^*_{\max}}^2 \log^2(p) - 24\ \ C^*_{\int}\ \mu^*_{\max}\ r^*_{\max}\ \mathfrak{s}\sqrt{\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}.$$

Remark 2.2 *Notice that, using (2.-5), the following relationship holds between $\|{\beta^*}_{T^*}\|_2^2$ and $\|\beta_T\|_2^2$ if all the coefficients β_j, $j \in \mathcal{J}_k$ have the same sign for all $k \in \mathcal{K}$:*

$$\|{\beta^*}_{T^*}\|_2^2 \geq \|\beta_T\|_2^2.$$

In this case, one can replace $\|{\beta^*}_{T^*}\|_2^2$ by $\|\beta_T\|_2^2$ in Assumption (2.9) and merge Assumption (2.8) and Assumption (2.9) by taking the maximum of their respective right hand side and obtain a simpler assumption.

Assumptions 2.10 *The support of $\beta^*_{T^*}$ is random and uniformly distributed among subsets of $\{1, \ldots, p\}$ with cardinal s^*. The sign of $\beta^*_{T^*}$ is random with uniform distribution on $\{-1, 1\}^{s^*}$.*

Remark 2.3 *This last assumption is a transposition to the proxy β^* of the conditions on β in [14].*

Main theorem

The main result of this paper is the following theorem.

Theorem 2.4 *Set $\lambda = 2\sigma\sqrt{2\alpha\ \log(p)}$. Assume that X is drawn from the Gaussian mixture model of Section 2 with $\mathcal{K}$ drawn uniformly at random among all possible index subsets of $\{1, \ldots, K\}$ with cardinal s^*. Let Assumptions 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8 and 2.9 hold. Then, we have*

$$\frac{1}{2}\|Xh\|_2^2 \leq s^* \frac{3}{2} r_*\ \lambda \left(\frac{3}{2}\ \lambda + \sqrt{1 + r_*}\ \delta \|\mathfrak{C}_T \beta_T\|_2\right) + \frac{1}{2}\delta^2 \|X\beta\|_2^2$$

with $r_* = 1.1 \cdot r \ \ (1.1 + 0.11 \cdot r)$ *and for any* δ *satisfying*

$$\begin{aligned}
\delta \ \geq \ & 4\mathfrak{s}\left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right)\left(1 + 8\sqrt{2}\ \sqrt{\alpha\log(p) + \log(2n+2)}\sqrt{s^*\rho_{\mathfrak{C}}}\right) \\
& + \left(12\ C_{\int}\ \mathfrak{s}\ \sqrt{n}\ r_{\max} + \alpha\log(p)\ \mu_{\max}\right)\ \sqrt{s^*\rho_{\mathfrak{C}}} \\
& + 4\mathfrak{s}\sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\left(1 + 2\sqrt{2}\ \sqrt{\rho_{\mathfrak{C}}}\sqrt{\alpha\log(p) + \log(2n+2))}\right)} \\
& + \left(24\ r^*_{\max}\ \mathfrak{s}\sqrt{\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\ C^*_{\int}} + \mu^*_{\max}\ \alpha\log(p)\right)\ \sqrt{\rho_{\mathfrak{C}}}, \quad (2.\text{-}18)
\end{aligned}$$

3 Proof of Theorem 3.1

Some parts of the proof closely follow the key arguments in the proof of [14, Theorem 1.2]. Their adaptation to the present setting is however sometimes nontrivial. We present all the details for the sake of completeness.

Preliminaries: Candès and Plan's conditions

The following proposition will be much used in the arguments.

Proposition 3.1 *We have the following properties:*

(I)

$$\mathbb{P}\left(\|\mathfrak{C}^t_{\mathcal{K}}\mathfrak{C}_{\mathcal{K}} - I_s\| \geq \frac{1}{2}\right) \ \leq \ \frac{216}{p^\alpha}. \quad (3.\text{-}17)$$

(II)

$$\mathbb{P}\left(\left\|X^t_{T^*}X_{T^*} - I\right\| \geq 1.1 \cdot r(1.1 + 0.11 \cdot r)\right) \ \leq \ \frac{219}{p^\alpha}. \quad (3.\text{-}16)$$

(III)

$$\mathbb{P}\left(\|X^t z\|_\infty \geq \sigma\sqrt{2\alpha\ \log(p)}\right) \ \leq \ \frac{1}{p^\alpha}. \quad (3.\text{-}15)$$

(IV)

$$\begin{aligned}
& \left\|X^t_{T^{*c}}X_{T^*}(X^t_{T^*}X_{T^*})^{-1}X^t_{T^*}z\right\|_\infty + \lambda\ \left\|X^t_{T^{*c}}X_{T^*}(X^t_{T^*}X_{T^*})^{-1}sign\,(\beta^*_{T^*})\right\|_\infty \\
& \leq \sigma\sqrt{1 + 1.1 \cdot r(1.1 + 0.11 \cdot r)} + \frac{1}{2}\ \lambda \quad (3.\text{-}15)
\end{aligned}$$

extbfProof. See Appendix 6. □

Controlling $\|X\beta - X\beta^*\|_2$ by $\|X\beta\|$

Proposition 3.2 *One has*

$$\mathbb{P}(\|X\beta - X\beta^*\|_2 \geq \delta \ \|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2) \quad \leq \quad \frac{1}{p^\alpha}$$

extbfProof. The proof is divided into four steps, for the sake of clarity.
Step 1. Let

$$\tilde{E}_T \quad = \quad X_T - \mathfrak{C}_{\mathcal{K}_T}.$$

where, since $\mathcal{K}_T$ is supposed to be a list with possible repetitions, the matrix $\mathfrak{C}_{\mathcal{K}_T}$ has correspondingly possible column repetitions, and

$$\tilde{E}^*_T \quad = \quad X_{T^*} - \mathfrak{C}_{\mathcal{K}_{T^*}}.$$

Thus, using (2.-6),

$$\|X\beta - X\beta^*\|_2 \quad = \quad \|\tilde{E}_T\beta_T - \tilde{E}^*_{T^*}\beta^*_{T^*}\|_2,$$

which, by the triangular inequality, gives

$$\|X\beta - X\beta^*\|_2 \quad \leq \quad \|\tilde{E}_T\beta_T\|_2 + \|\tilde{E}^*_{T^*}\beta^*_{T^*}\|_2.$$

Step 2: Control of $\|\tilde{E}_T\beta_T\|_2$. The column $j \in T$ of the matrix $\tilde{E}_T$ has the expression

$$\tilde{E}_j \quad = \quad \frac{\mathfrak{C}_{k_j} + E_j}{\|\mathfrak{C}_{k_j} + E_j\|_2} - \mathfrak{C}_{k_j}.$$

We may decompose the quantity $\|\tilde{E}_T\beta_T\|_2$ as

$$\|\tilde{E}_T\beta_T\|_2 \quad = \quad \|A\|_2 + \|B\|_2,$$

where

$$A \quad = \quad \sum_{j\in T}\left(\frac{1}{\|\mathfrak{C}_{k_j} + E_j\|_2} - 1\right)\mathfrak{C}_{k_j}\beta_j$$

and

$$B \quad = \quad \sum_{j\in T}\frac{1}{\|\mathfrak{C}_{k_j} + E_j\|_2}E_j\beta_j.$$

We have the following bound for A.

Lemma 3.3

$$\mathbb{P}\left(\|A\|_2 \geq 4\mathfrak{s}\ \left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right)\left(1 + 8\sqrt{2}\ \sqrt{\alpha\log(p) + \log(2n+2)}\sqrt{s^*\rho_{\mathfrak{C}}}\right)\|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2\right) \leq \frac{C+1}{p^\alpha}. \tag{3.-24}$$

extbfProof. See Appendix 4. □ Turning to B, we have the following result.

Lemma 3.4 *We have*

$$\mathbb{P}\left(\|B\|_2 \geq \left(12\ C_{\int}\ \mathfrak{s}\ \sqrt{n}\ r_{\max} + \alpha \log(p)\ \mu_{\max} \right)\ \sqrt{s^* \rho_{\mathfrak{C}}} \|\mathfrak{C}_{\mathcal{K}_T} \beta_T\|_2 \right) \leq \frac{2}{p^{\alpha}}.$$

extbfProof. See Appendix 4. □

Step 3: Control of $\|\tilde{E}^*_{T^*} \beta^*_{T^*}\|_2$. The column $j^* \in T^*$ of the matrix $\tilde{E}^*_{T^*}$ has the expression

$$\tilde{E}^*_{j^*} \quad = \quad \frac{\mathfrak{C}_{k_{j^*}} + E_{j^*}}{\|\mathfrak{C}_{k_{j^*}} + E_{j^*}\|_2} - \mathfrak{C}_{k_{j^*}}.$$

We will procede as in Step 2. Define

$$W^*_{j^*} \quad = \quad \frac{1}{\|\mathfrak{C}_{k_{j^*}} + E_{j^*}\|_2} - 1.$$

Notice that $\tilde{E}^*_{j^*}$ can be written

$$\tilde{E}^*_{T^*} \beta^*_{T^*} \quad = \quad A^* + B^*$$

with

$$A^* = \quad \sum_{j^* \in T^*} W^*_{j^*} \beta^*_{j^*} A^*_{j^*},$$

where

$$A^*_{j^*} = \quad \begin{bmatrix} 0 & \mathfrak{C}^t_{k_{j^*}} \\ \mathfrak{C}^t_{k_{j^*}} & 0 \end{bmatrix}.$$

and

$$B^* = \quad \sum_{j^* \in T^*} \beta^*_{j^*} B^*_{j^*}, \ \text{ where } B_{j^*} = \quad \frac{E_{j^*}}{\|\mathfrak{C}_{k_{j^*}} + E_{j^*}\|_2}.$$

We begin with the study of A^*.

Lemma 3.5 *We have*

$$\mathbb{P}\Bigg(\|A^*\|_2 \geq 4\mathfrak{s} \sqrt{n\ \left(\frac{\alpha\ (1 - e^{-1})}{\vartheta_*\ C_{\chi}} \right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu - 1}} \right)^{\frac{1}{n}}} \left(1 + 2\sqrt{2}\ \sqrt{\rho_{\mathfrak{C}}} \sqrt{\alpha \log(p) + \log(2n + 2))} \right) \|\mathfrak{C}_{\mathcal{K}_T} \beta_T\|_2 \Bigg) \leq \frac{2}{p^{\alpha}}.$$

extbfProof. See Appendix 4. □ Turning to B^*, we have the following result.

Lemma 3.6 *We have*

$$\mathbb{P}\Bigg(\|B^*\| \geq \Bigg(24\ r^*_{\max}\ \mathfrak{s} \sqrt{\left(\frac{\alpha\ (1 - e^{-1})}{\vartheta_*\ C_{\chi}} \right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu - 1}} \right)^{\frac{1}{n}}}\ C^*_{\int} + \mu^*_{\max}\ \alpha \log(p) \Bigg)\ \sqrt{\rho_{\mathfrak{C}}}\ \|\mathfrak{C}_T \beta_T\|_2 \Bigg) \leq \frac{2}{p^{\alpha}}.$$

extbfProof. See Appendix 4. □

Step 4: Conclusion. Combining Lemmæ 1.1, 3.4, 3.5 and 3.6, we obtain that for any δ such that (2.-15) we have

$$\mathbb{P}(\|X\beta - X\beta^*\|_2 \geq \delta \|\mathfrak{C}_T \beta_T\|_2) \quad \leq \quad \frac{1}{p^\alpha}.$$

□

The prediction bound

By definition, the LASSO estimator satisfies

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \quad \leq \quad \frac{1}{2}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1. \tag{3.-36}$$

One may introduce $X\beta$ in this expression and obtain

$$\frac{1}{2}\|y - X\beta + X(\beta - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}\|_1 \quad \leq \quad \frac{1}{2}\|y - X\beta + X(\beta - \beta^*)\|_2^2 + \lambda\|\beta^*\|_1,$$

from which we deduce

$$\begin{aligned} \frac{1}{2}\|X(\beta - \hat{\beta})\|_2^2 \quad \leq \quad & \langle y - X\beta, X(\hat{\beta} - \beta^*)\rangle \\ & -\lambda\left(\|\hat{\beta}\|_1 - \|\beta^*\|_1\right) + \frac{1}{2}\|X(\beta - \beta^*)\|_2^2. \end{aligned} \tag{3.-36}$$

Set $h^* := \hat{\beta} - \beta^*$. Using sparsity of β^*, we obtain that ${h^*}_{T^{*c}} = \hat{\beta}_{T^{*c}} - {\beta^*}_{T^{*c}} = \hat{\beta}_{T^{*c}}$. Thus, we have

$$\begin{aligned} \|\hat{\beta}\|_1 - \|\beta^*\|_1 \quad &= \quad \|\beta^* + h\|_1 - \|\beta^*\|_1 \\ &= \quad \|{\beta^*}_{T^*} + {h^*}_{T^*}\|_1 + \|{\beta^*}_{T^{*c}} + {h^*}_{T^{*c}}\|_1 - \|{\beta^*}_{T^*}\|_1 \\ &= \quad \|{\beta^*}_{T^*} + {h^*}_{T^*}\|_1 - \|{\beta^*}_{T^*}\|_1 + \|{h^*}_{T^{*c}}\|_1. \end{aligned}$$

Since, for any b with no zero component, the gradient of $\|\cdot\|_1$ at b is $sign(b)$, the subgradient inequality gives

$$\|{\beta^*}_{T^*} + {h^*}_{T^*}\|_1 \quad \geq \quad \|{\beta^*}_{T^*}\|_1 + \langle sign\left({\beta^*}_{T^*}\right), {h^*}_{T^*}\rangle$$

and combining this latter inequality with (3.-36), we obtain

$$\begin{aligned} \frac{1}{2}\|X(\beta - \hat{\beta})\|_2^2 \quad \leq \quad & \langle y - X\beta, Xh^*\rangle - \lambda\langle sign\left({\beta^*}_{T^*}\right), {h^*}_{T^*}\rangle \\ & -\lambda\|{h^*}_{T^{*c}}\|_1 + \frac{1}{2}\|X(\beta - \beta^*)\|_2^2. \end{aligned} \tag{3.-40}$$

Set $r := \beta^* - \beta$ and $h := \hat{\beta} - \beta$. Using these notations, equation (3.-40) may be written

$$\begin{aligned} \frac{1}{2}\|Xh\|_2^2 \quad \leq \quad & \langle z, Xh^*\rangle - \lambda\langle sign\left({\beta^*}_{T^*}\right), {h^*}_{T^*}\rangle \\ & -\lambda\|{h^*}_{T^{*c}}\|_1 + \frac{1}{2}\|Xr\|_2^2. \end{aligned} \tag{3.-40}$$

Using the fact that

$$\langle X^t z, h^*\rangle \quad = \quad \langle X^t_{T^*} z, {h^*}_{T^*}\rangle + \langle X^t_{T^{*c}} z, {h^*}_{T^{*c}}\rangle$$

and the following majorization based on (3.-15)

$$\begin{aligned}\langle X^t_{T^{*c}} z, {h^*}_{T^{*c}}\rangle &\leq \|{h^*}_{T^{*c}}\|_1 \|X'_{T^{*c}} z\|_\infty \\ &\leq \frac{1}{2}\,\lambda\,\|{h^*}_{T^{*c}}\|_1,\end{aligned}$$

we obtain that

$$\frac{1}{2}\|Xh\|_2^2 \leq \langle v, {h^*}_{T^*}\rangle - (1-\frac{1}{2})\,\lambda\|{h^*}_{T^{*c}}\|_1 + \frac{1}{2}\|Xr\|_2^2,$$

where $v := X^t_{T^*} z - \lambda\, sign\,({\beta^*}_{T^*})$.

Now, observe that

$$\begin{aligned}\langle v, {h^*}_{T^*}\rangle &= \langle v, (X^t_{T^*}X_{T^*})^{-1}X^t_{T^*}X_{T^*}{h^*}_{T^*}\rangle \\ &= \langle (X_{T^*}X_{T^*})^{-1}v, X^t_{T^*}X_{T^*}{h^*}_{T^*}\rangle \\ &= \underbrace{\langle (X^t_{T^*}X_{T^*})^{-1}v, X^t_{T^*}Xh^*\rangle}_{A_1} - \underbrace{\langle (X^t_{T^*}X_{T^*})^{-1}v, X^t_{T^*}X_{T^{*c}}{h^*}_{T^{*c}}\rangle}_{A_2}.\end{aligned}$$

Let us begin by studying A_2. We have that

$$\begin{aligned}A_2 &\geq -\|X^t_{T^{*c}}X_{T^*}(X^t_{T^*}X_{T^*})^{-1}v\|_\infty\|h_{T^{*c}}\|_1 \\ &\geq -\|X^t_{T^{*c}}X_{T^*}(X^t_{T^*}X_{T^*})^{-1}X^t_{T^*}z\|_\infty\|h_{T^{*c}}\|_1 \\ &\quad -\lambda\,\|X^t_{T^{*c}}X_{T^*}(X^t_{T^*}X_{T^*})^{-1}sign\,({\beta^*}_{T^*})\,\|_\infty\|{h^*}_{T^{*c}}\|_1 \\ &\geq -\left(\sigma\sqrt{1+1.1\cdot r\ (1.1+0.11\cdot r)} + \frac{1}{2}\,\lambda\right)\,\|{h^*}_{T^{*c}}\|_1\end{aligned}$$

by (3.-15). Thus

$$\langle v, {h^*}_{T^*}\rangle \leq A_1 + \left(\sigma\sqrt{1+1.1\cdot r\ (1.1+0.11\cdot r)} + \frac{1}{2}\,\lambda\right)\,\|h_{T^{*c}}\|_1$$

and since, by (2.-8),

$$\left(\sigma\sqrt{1+1.1\cdot r\ (1.1+0.11\cdot r)} + \frac{1}{2}\,\lambda\right) \leq \lambda,$$

we deduce that

$$\frac{1}{2}\|Xh\|_2^2 \leq A_1 + \frac{1}{2}\|Xr\|_2^2,$$

Let us now bound A_1 from above. We have that

$$A_1 \leq \underbrace{\|X^t_{T^*}Xh^*\|_\infty}_{B_1}\underbrace{\|(X^t_{T^*}X_{T^*})^{-1}v\|_1}_{B_2}$$

Firstly,

$$\begin{aligned}B_1 &\leq \|X^t_{T^*}(X\beta^* - y)\|_\infty + \|X^t_{T^*}(X\hat{\beta} - y)\|_\infty \\ &\leq \|X^t_{T^*}(Xr - z)\|_\infty + \|X^t_{T^*}(y - X\hat{\beta})\|_\infty \\ &\leq \frac{1}{2}\lambda + \|X^t_{T^*}Xr\|_\infty + \lambda\end{aligned}$$

where we used (3.-15), and the optimality condition for the LASSO estimator. Secondly,

$$\begin{aligned} B_2 &\leq \sqrt{s^*}\|(X_{T^*}^t X_{T^*})^{-1}v\|_2 \\ &\leq \sqrt{s^*}\|(X_{T^*}^t X_{T^*})^{-1}\|\|v\|_2 \\ &\leq s^*\|(X_{T^*}^t X_{T^*})^{-1}\|\|v\|_\infty. \end{aligned}$$

Moreover, (3.-16) gives $\|(X_{T^*}^t X_{T^*})^{-1}\| \leq 1.1 \cdot r \ (1.1 + 0.11 \cdot r)$ and

$$\|v\|_\infty \leq \|X_{T^*}^t z\|_\infty + \lambda \leq \frac{3}{2}\,\lambda$$

Thus, we obtain that

$$A_1 \leq s^* 1.1 \cdot r \ (1.1 + 0.11 \cdot r)\,\frac{3}{2}\,\lambda\left(\frac{3}{2}\,\lambda + \|X_{T^*}^t X r\|_\infty\right)$$

and thus,

$$\frac{1}{2}\|Xh\|_2^2 \leq s^* 1.1 \cdot r \ (1.1 + 0.11 \cdot r)\,\frac{3}{2}\,\lambda\left(\frac{3}{2}\,\lambda + \|X_{T^*}^t X r\|_\infty\right) + \frac{1}{2}\|Xr\|_2^2.$$

Since $\|X_{T^*}^t X r\|_\infty \leq \|X_{T^*}^t X r\|_2$ and since $\|X_{T^*}^t X r\|_2 \leq \sqrt{1 + 1.1 \cdot r \ (1.1 + 0.11 \cdot r)}\|Xr\|_2$, we obtain

$$\frac{1}{2}\|Xh\|_2^2 \leq s^* 1.1 \cdot r \ (1.1 + 0.11 \cdot r)\,\frac{3}{2}\,\lambda\left(\frac{3}{2}\,\lambda + \sqrt{1 + 1.1 \cdot r \ (1.1 + 0.11 \cdot r)}\|Xr\|_2\right) + \frac{1}{2}\|Xr\|_2^2.$$

Moreover, Proposition 3.2 yields

$$\frac{1}{2}\|Xh\|_2^2 \leq s^* 1.1 \cdot r \ (1.1 + 0.11 \cdot r)\,\frac{3}{2}\,\lambda\left(\frac{3}{2}\,\lambda + \sqrt{1 + 1.1 \cdot r \ (1.1 + 0.11 \cdot r)}\ \delta\|\mathfrak{C}_T \beta_T\|_2\right) + \frac{1}{2}\delta^2\|X\beta\|_2^2$$

which completes the proof.

4 Technical lemmæ

Proof of Lemma 1.1

We have that

$$\|\mathfrak{C}_{k_j}\|_2 - \|E_j\|_2 \leq \|\mathfrak{C}_{k_j} + E_j\|_2 \leq \|\mathfrak{C}_{k_j}\|_2 + \|E_j\|_2.$$

Moreover, since $\|E_j\|_2^2/\mathfrak{s}^2$ follows the χ_n^2-distribution, the scalar Chernov bound gives

$$\mathbb{P}\left(\left|\frac{\|E_j\|_2}{\mathfrak{s}} - \sqrt{n}\right| \geq u\right) \leq C\exp\left(-cu^2\right) \tag{4.-68}$$

for some constants c and C. Let W_j denote the following variable.

$$W_j = \frac{1}{\|\mathfrak{C}_{k_j} + E_j\|_2} - 1,$$

and let $\mathcal{E}_\alpha$ denote the event

$$\begin{aligned}\mathcal{E}_\alpha &= \cap_{j\in T}\Bigg\{ -\frac{\mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)}{1+\mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)} \leq \frac{1}{\|\mathfrak{C}_{k_j}+E_j\|_2}-1 \\ &\qquad \leq \frac{\mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)}{1-\mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)}\Bigg\}.\end{aligned}$$

Taking $u=\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}$, we obtain that

$$\mathbb{P}\left(\mathcal{E}_\alpha\right) \geq 1-\frac{C}{p^\alpha}.$$

On the other hand, we can write $\|A\|_2$ as

$$\|A\|_2 = \Big\|\sum_{j\in T} A_j\Big\|,$$

where A_j is the matrix

$$A_j = W_j\begin{bmatrix} 0 & \mathfrak{C}_{k_j}^t\beta_j \\ \mathfrak{C}_{k_j}\beta_j & 0\end{bmatrix}$$

Thus, by the triangular inequality, we have

$$\|A\|_2 \leq \Big\|\sum_{j\in T} A_j-\mathbb{E}\left[A_j\mid\mathcal{E}_\alpha\right]\Big\|+\Big\|\sum_{j\in T}\mathbb{E}\left[A_j\mid\mathcal{E}_\alpha\right]\Big\|, \tag{4.-73}$$

and we may apply the Matrix Hoeffding inequality recalled in Appendix 5. We have that

$$\|A_j\| = |W_j||\beta_j|$$

which implies that, on $\mathcal{E}_\alpha$, we have

$$\|A_j-\mathbb{E}\left[A_j\mid\mathcal{E}_\alpha\right]\| \leq 2\frac{\mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)}{1-\mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)}|\beta_j|,$$

which, by Assumption 2.7, gives that

$$\|A_j-\mathbb{E}\left[A_j\mid\mathcal{E}_\alpha\right]\| \leq 4\mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)|\beta_j|.$$

The Matrix Hoeffding inequality, first applied to the sum and then to its opposite, yields

$$\begin{aligned}&\mathbb{P}\left(\|\sum_{j\in T}A_j-\mathbb{E}\left[A_j\mid\mathcal{E}_\alpha\right]\|_2\geq u\mid\mathcal{E}_\alpha\right) \\ &\qquad\leq 2(n+1)\cdot\exp\left(-\frac{u^2}{8\cdot 16\,\mathfrak{s}^2\left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)^2\|\beta_T\|_2^2}\right)\end{aligned} \tag{4.-76}$$

On the other hand, we have that

$$\left\| \sum_{j\in T} \mathbb{E}\left[A_j \mid \mathcal{E}_\alpha\right] \right\| \quad = \quad \left\| \sum_{j\in T} \mathbb{E}\left[W_j \mid \mathcal{E}_\alpha\right] \begin{bmatrix} 0 & \mathfrak{C}_{k_j}^t \beta_j \\ \mathfrak{C}_{k_j}\beta_j & 0 \end{bmatrix} \right\|.$$

Notice that, since $\|\mathfrak{C}_{k_j}\|_2 = 1$, then, $\|\mathfrak{C}_{k_j} + E_j\|_2^2/\mathfrak{s}^2$ has a noncentral-χ^2 distribution with non-centrality parameter equal to $1/\mathfrak{s}^2$, for all $j \in T$. Thus, we deduce that all the variables $\|\mathfrak{C}_{k_j} + E_j\|_2^2$, $j \in T$, have the same distribution and in particular, the same conditional expectation given $\mathcal{E}_\alpha$. Therefore,

$$\begin{aligned} \left\| \sum_{j\in T} \mathbb{E}\left[A_j \mid \mathcal{E}_\alpha\right] \right\| &= \left|\mathbb{E}\left[W_1 \mid \mathcal{E}_\alpha\right]\right| \left\| \sum_{j\in T} \begin{bmatrix} 0 & \mathfrak{C}_{k_j}^t \beta_j \\ \mathfrak{C}_{k_j}\beta_j & 0 \end{bmatrix} \right\| \\ &= \left|\mathbb{E}\left[W_1 \mid \mathcal{E}_\alpha\right]\right| \|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2. \end{aligned}$$

But since

$$\left|\mathbb{E}\left[W_1 \mid \mathcal{E}_\alpha\right]\right| \quad \leq \quad 4\mathfrak{s}\left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right),$$

we obtain

$$\left\| \sum_{j\in T} \mathbb{E}\left[A_j \mid \mathcal{E}_\alpha\right] \right\| \quad = \quad 4\mathfrak{s}\left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right) \|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2.$$

Combining this latter inequality with (4.-76), (4.-73) becomes

$$\begin{aligned} &\mathbb{P}\left(\|A\|_2 \geq 4\mathfrak{s}\left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right)\|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2 + u \mid \mathcal{E}_\alpha\right) \\ &\qquad \leq 2(n+1)\cdot\exp\left(-\frac{u^2}{8\cdot 16\ \mathfrak{s}^2\ \left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right)^2 \|\beta_T\|_2^2}\right). \end{aligned}$$

Since, for any event $\mathcal{A}$,

$$\mathbb{P}\left(\mathcal{A}\right) \quad \leq \quad \mathbb{P}\left(\mathcal{A} \mid \mathcal{E}_\alpha\right) + \mathbb{P}\left(\mathcal{E}_\alpha^c\right),$$

we obtain that

$$\begin{aligned} &\mathbb{P}\left(\|A\|_2 \geq 4\mathfrak{s}\left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right)\|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2 + u\right) \\ &\qquad \leq 2(n+1)\cdot\exp\left(-\frac{u^2}{8\cdot 16\ \mathfrak{s}^2\ \left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right)^2 \|\beta_T\|_2^2}\right) + \frac{C}{p^\alpha}. \end{aligned}$$

Let us now choose u such that

$$2(n+1)\cdot\exp\left(-\frac{u^2}{8\cdot 16\ \mathfrak{s}^2\ \left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right)^2 \|\beta_T\|_2^2}\right) \quad = \quad \frac{1}{p^\alpha},$$

i.e.

$$u \quad = \quad 8\sqrt{2}\ \mathfrak{s}\ \left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)\|\beta_T\|_2\sqrt{\alpha\log(p)+\log(2n+2))}.$$

Therefore, we obtain that

$$\begin{aligned}\mathbb{P}\Bigg(\|A\|_2 \geq 4\mathfrak{s}\ \left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)\Bigg(&\|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2 \\ &+8\sqrt{2}\ \sqrt{\alpha\log(p)+\log(2n+2)}\|\beta_T\|_2\Bigg)\Bigg) \\ &\leq \frac{C+1}{p^\alpha}. \end{aligned} \tag{4.-89}$$

Recall that we assumed the β_j associated to the same cluster to have the same sign. Thus, we obtain that

$$\|\beta_T\|_1 \quad = \|\beta^*_{T^*}\|_1 \leq \quad \sqrt{s^*}\|\beta^*_{T^*}\|_2,$$

and using the version of the Invertibility Condition for $\mathfrak{C}$ (3.-17), we get

$$\|\beta_T\|_1 \quad = \quad \sqrt{s^*\rho_{\mathfrak{C}}}\ \|\mathfrak{C}_{\mathcal{K}^*_T}\beta^*_{T^*}\|_2,$$

and thus,

$$\|\beta_T\|_2 \quad = \quad \sqrt{s^*\rho_{\mathfrak{C}}}\ \|\mathfrak{C}_{\mathcal{K}^*_T}\beta^*_{T^*}\|_2$$

and, using the definition of β_{T^*},

$$\|\beta_T\|_2 \quad = \quad \sqrt{s^*\rho_{\mathfrak{C}}}\ \|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2. \tag{4.-91}$$

Thus, (4.-87) gives

$$\begin{aligned}\mathbb{P}\left(\|A\|_2 \geq 4\mathfrak{s}\ \left(\sqrt{n}+\sqrt{\frac{\alpha}{c}\log(p)+\frac{1}{c}\log(s)}\right)\left(1+8\sqrt{2}\ \sqrt{\alpha\log(p)+\log(2n+2)}\sqrt{s^*\rho_{\mathfrak{C}}}\right)\|\mathfrak{C}_{\mathcal{K}_T}\beta_T\right. \\ \left.\leq \frac{C+1}{p^\alpha}.\right. \end{aligned} \tag{4.}$$

Proof of Lemma 3.4

Recall that

$$\|B\|_2 \quad = \quad \Big\|\sum_{j\in T}\frac{\beta_j}{\|\mathfrak{C}_{k_j}+E_j\|_2}E_j\Big\|_2.$$

Will be use Talagrand's concentration inequality and Dudley's entropy integral bound to study $\|B\|_2$. We start with some preliminary results.

Preliminaries

Let us define the following event:

$$\mathcal{F}_\alpha \quad = \quad \mathcal{E}_\alpha \cap \left\{ \left\| E_T^t \right\| \leq \mathfrak{s} \left(\sqrt{n} + \sqrt{s} + \sqrt{2\alpha \log(p)} \right) \right\}.$$

Since E_T^t is i.i.d. with Gaussian entrees $\mathcal{N}(0, \mathfrak{s})$, Theorem 5.5 in Appendix 5 gives

$$\mathbb{P}\left(\left\| E_T^t \right\| \geq \mathfrak{s} \left(\sqrt{n} + \sqrt{s} + \sqrt{2\alpha \log(p)} \right) \right) \quad \leq \quad \frac{2}{p^\alpha}.$$

Thus, the union bound gives that $\mathbb{P}(\mathcal{F}_\alpha) \geq 3/p^\alpha$. Let us now turn to the task of bounding $\|B\|_2$.

Concentration of $\|B\|_2$ using Talagrand's inequality

Notice that on $\mathcal{F}_\alpha$, we have

$$\left\| \sum_{j \in T} \frac{\beta_j}{\|\mathfrak{C}_{k_j} + E_j\|_2} E_j \right\|_2 \quad \leq \quad \max_b \left\| \sum_{j \in T} \frac{\beta_j}{b} E_j \right\|_2,$$

where the maximum is over all

$$b \in \left[1 - \mathfrak{s} \left(\sqrt{n} + \sqrt{\frac{\alpha}{c} \log(p) + \frac{1}{c} \log(s)} \right), 1 + \mathfrak{s} \left(\sqrt{n} + \sqrt{\frac{\alpha}{c} \log(p) + \frac{1}{c} \log(s)} \right) \right].$$

Thus, on $\mathcal{F}_\alpha$,

$$\|B\|_2 \quad \leq \quad \max_{b, \|w\|_2 = 1} \langle w, \sum_{j \in T} \frac{\beta_j}{b} E_{i,j} \rangle,$$

the main advantage of this former inequality being that of involving the supremum of a simple Gaussian process. Now, we have

$$\mathbb{P}\left(\|B\|_2 - \mathbb{E}\left[\max_{b, \|w\|_2 = 1} \langle w, \sum_{j \in T} \frac{\beta_j}{b} E_j \rangle \mid \mathcal{F}_\alpha \right] \geq u \mid \mathcal{F}_\alpha \right)$$
$$\leq \mathbb{P}\left(\max_{b, \|w\|_2 = 1} \langle w, \sum_{j \in T} \frac{\beta_j}{b} E_j \rangle - \mathbb{E}\left[\max_{b, \|w\|_2 = 1} \langle w, \sum_{j \in T} \frac{\beta_j}{b} E_j \rangle \mid \mathcal{F}_\alpha \right] \geq u \mid \mathcal{F}_\alpha \right).$$

Let

$$M_{b,w} \quad = \quad \langle w, \sum_{j \in T} \frac{\beta_j}{b} E_j \rangle.$$

In order to apply Talagrand's concentration inequality, we have to bound the $M_{b,w}$ on $\mathcal{E}_\alpha$, and its conditional variance given $\mathcal{F}_\alpha$. First, by the Cauchy-Schwartz inequality, we have

$$M_{b,w} \quad \leq \quad \frac{1}{b} \|\beta_T\|_2 \sqrt{\sum_{j \in T} (w^t E_j)^2},$$

and thus,

$$\begin{aligned} M_{b,w} &\leq \frac{1}{b}\|\beta_T\|_2\|E_T^t w\|_2 \\ &\leq \frac{1}{b}\|\beta_T\|_2\|E_T^t\|\|w\|_2. \end{aligned}$$

Thus, on $\mathcal{F}_\alpha$, using the fact that $\|w\|_2 = 1$, we have

$$M_{b,w} \leq \mu_{\max}\,\|\beta_T\|_2,$$

where $\mu_{\max}$ is given by (2.-7). Let us now turn to the conditional variance of $M_{b,w}$ given $\mathcal{F}_\alpha$. We have

$$Var\left(M_{b,w} \mid \mathcal{F}_\alpha\right) = \sum_{j\in T}\frac{\beta_j}{b}Var\left(E_j^t w \mid \mathcal{F}_\alpha\right),$$

and, using the Cauchy-Schwartz inequality again, we obtain

$$Var\left(M_{b,w} \mid \mathcal{F}_\alpha\right) = \frac{\|\beta_T\|_2}{b}\sqrt{\sum_{j\in T}Var^2\left(E_j^t w \mid \mathcal{F}_\alpha\right)}.$$

On the other hand, notice that, conditionally on $\mathcal{F}_\alpha$, $E_j^t w$ is centered. This can easily be seen from the invariance of both the Gaussian law and the event $\mathcal{F}_\alpha$ under the action of orthogonal transformations. Therefore, we have

$$Var\left(E_j^t w\right) \geq Var\left(E_j^t w \mid \mathcal{F}_\alpha\right)\left(1 - \frac{3}{p^\alpha}\right).$$

Moreover, using the fact that $\|w\|_2 = 1$,

$$Var\left(E_j^t w\right) = \mathfrak{s}^2.$$

Therefore,

$$Var\left(M_{b,w} \mid \mathcal{F}_\alpha\right) = \frac{\sqrt{s}\,\mathfrak{s}^2}{b}\|\beta_T\|_2.$$

Using the lower bound on b, we finally obtain

$$Var\left(M_{b,w} \mid \mathcal{F}_\alpha\right) \leq \sigma_{\max}^2\,\|\beta_T\|_2,$$

where $\sigma_{\max}^2$ is defined by (2.-6). With the bound on $M_{b,w}$ and its conditional variance in hand, we are ready to use Talagrand's concentration inequality recalled in Appendix 5. Thus, Theorem 5.6 gives

$$\mathbb{P}\left(\max_{b,\|w\|_2=1}\frac{M_{b,w}}{\mu_{\max}\,\|\beta_T\|_2} \geq \mathbb{E}\left[\max_{b,\|w\|_2=1}\frac{M_{b,w}}{\mu_{\max}\,\|\beta_T\|_2} \mid \mathcal{F}_\alpha\right] + \sqrt{2u\gamma} + \frac{u}{3} \mid \mathcal{F}_\alpha\right) \tag{4.-110}$$

$$\leq \exp\left(-u\right), \tag{4.-109}$$

with

$$\gamma = n\frac{\sigma_{\max}^2}{\mu_{\max}^2\,\|\beta_T\|_2^2} + \mathbb{E}\left[\max_{b,\|w\|_2=1}\frac{M_{b,w}}{\mu_{\max}\,\|\beta_T\|_2} \mid \mathcal{F}_\alpha\right].$$

Control of the conditional expectation of $\max_{b,\|w\|_2=1} \frac{M_{b,w}}{\mu_{\max}}$

Notice that

$$\mathbb{E}\left[\max_{b,\|w\|_2=1}\langle w, \sum_{j\in T}\frac{\beta_j}{b}E_j\rangle\right] \geq \mathbb{E}\left[\max_{b,\|w\|_2=1}\langle w, \sum_{j\in T}\frac{\beta_j}{b}E_j\rangle \mid \mathcal{F}_\alpha\right]\left(1-\frac{1}{p^\alpha}\right).$$

Therefore, our task boils down to controling the supremum of a centered gaussian process. For this purpose, let $\tilde{w} = w/b$, which implies that

$$\mathbb{E}\left[\max_{b,\|w\|_2=1}\langle w, \sum_{j\in T}\frac{\beta_j}{b}E_j\rangle\right] = \mathbb{E}\left[\max_{\tilde{w}\in\mathcal{T}}\langle \tilde{w}, \sum_{j\in T}\beta_j E_j\rangle\right]$$

where $\mathcal{T}$ denotes the spherical shell between the sphere centered at zero with radius $r_{\max} = 1 + \mathfrak{s}\left(\sqrt{n} + \sqrt{\frac{\alpha}{c}\log(p) + \frac{1}{c}\log(s)}\right)$ and the sphere centered at zero with radius $r_{\min} = 2 - r_{\max}$. This can of course be performed using Dudley's entropy bound recalled in Section 5. In the terminology of Section 5, the semi-metric d given by

$$d^2(\tilde{w}, \tilde{w}')) = \mathbb{E}\left[\left(\langle \tilde{w} - \tilde{w}', \sum_{j\in T}\beta_j E_j\rangle\right)^2\right].$$

The variables $\beta_j\ (w-w')^t E_j$, $j \in T$, are centered and have variance equal to $\mathfrak{s}^2\beta_j^2\|w-w'\|_2^2$. Thus,

$$d(w, w') = \mathfrak{s}\ \|\beta_T\|_2\ \|w - w'\|_2.$$

Let us now consider the entropy. An upper bound on the covering number of $\mathcal{T}$ with respect to the euclidean distance is given by

$$H(\varepsilon, \mathcal{T}) \leq n\log\left(\frac{3\ \mathfrak{s}\ r_{\max}\ \|\beta_T\|_2}{\varepsilon}\right).$$

Therefore, by Theorem 5.7, we obtain that

$$\mathbb{E}\left[\max_{\tilde{w}\in\mathcal{T}}\langle \tilde{w}, \sum_{j\in T}\beta_j E_j\rangle\right] \leq 12\sqrt{n}\int_0^{\sigma_G}\sqrt{\log\left(\frac{3\ \mathfrak{s}\ r_{\max}\|\beta_T\|_2}{\varepsilon}\right)}d\varepsilon,$$

with

$$\sigma_G = \mathfrak{s}\ r_{\max}\ \|\beta_T\|_2.$$

Using the change of variable $\varepsilon' = \frac{\varepsilon}{\mathfrak{s}\ r_{\max}\ \|\beta_T\|_2}$, we obtain

$$\mathbb{E}\left[\max_{\tilde{w}\in\mathcal{T}}\langle \tilde{w}, \sum_{j\in T}\beta_j E_j\rangle\right] \leq 12\ C_\int\ \mathfrak{s}\ r_{\max}\ \|\beta_T\|_2\sqrt{n}, \tag{4.-116}$$

where we recall that

$$C_\int = \int_0^3\sqrt{\log\left(\frac{3}{\varepsilon'}\right)}\ d\varepsilon'.$$

Conclusion of the proof

To sum up, combining (4.-109) and (4.-116)

$$\mathbb{P}\left(\|B\|_2 \geq 12\ C_{\int}\ \mathfrak{s}\ \sqrt{n}\ r_{\max}\ \|\beta_T\|_2 + \mu_{\max}\ \|\beta_T\|_2\ \left(\sqrt{2u\gamma} + \frac{u}{3}\right) \mid \mathcal{F}_\alpha\right) \quad \leq \quad \exp\left(-u\right) \tag{4.-116}$$

with

$$\gamma \leq n\ \frac{\sigma_{\max}^2}{\mu_{\max}^2\ \|\beta_T\|_2^2} + 12\ \frac{C_{\int}}{\mu_{\max}}\ \mathfrak{s}\ \sqrt{n}\ r_{\max}.$$

Thus,

$$\mathbb{P}\left(\|B\|_2 \geq \left(12\ C_{\int}\ \mathfrak{s}\ \sqrt{n}\ r_{\max} + \sqrt{2n\sigma_{\max}^2\ \frac{u}{\|\beta_T\|_2} + 12\ C_{\int}\mu_{\max}\ \mathfrak{s}\sqrt{n}\ r_{\max}} + \mu_{\max}\frac{u}{3}\right)\ \|\beta_T\|_2 \mid \mathcal{F}_\alpha\right)$$

$$\leq \exp\left(-u\right). \tag{4.-11}$$

Taking $u = \alpha \log(p)$ and using Assumption 2.8 gives

$$\mathbb{P}\left(\|B\|_2 \geq \left(12\ C_{\int}\ \mathfrak{s}\ \sqrt{n}\ r_{\max} + \alpha \log(p)\ \mu_{\max}\right)\ \|\beta_T\|_2 \mid \mathcal{F}_\alpha\right) \leq \frac{1}{p^\alpha}.$$

Using the same trick as before, we have

$$\mathbb{P}\left(\|B\|_2 \geq \left(12\ C_{\int}\ \mathfrak{s}\ \sqrt{n}\ r_{\max} + \alpha \log(p)\ \mu_{\max}\right)\ \|\beta_T\|_2\right) \leq \frac{2}{p^\alpha}.$$

Finally, using (4.-91), we have

$$\mathbb{P}\left(\|B\|_2 \geq \left(12\ C_{\int}\ \mathfrak{s}\ \sqrt{n}\ r_{\max} + \alpha \log(p)\ \mu_{\max}\right)\ \sqrt{s^* \rho_{\mathfrak{C}}}\|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2\right) \leq \frac{2}{p^\alpha}.$$

Proof of Lemma 3.5

We will use the same arguments based on the Matrix Hoeffding inequality as in 4. For this purpose, define

$$W_{j^*}^* \quad = \quad \frac{1}{\|\mathfrak{C}_{k_{j^*}} + E_{j^*}\|_2} - 1$$

and wrFite

$$\|A^*\|_2 \quad = \quad \left\|\sum_{j^* \in T^*} W_{j^*}^* \beta_{j^*}^* A_{j^*}^*\right\|.$$

We will need the following lemma.

Lemma 4.1 *Let*

$$\mathcal{E}_\alpha^* \quad = \quad \cap_{j^* \in T^*} \left\{\|E_{j^*}\|_2 \leq \mathfrak{s}\ \sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\right\}.$$

Then, $\mathbb{P}\left(\mathcal{E}_\alpha^*\right) \geq 1 - p^{-\alpha}$.

extbfProof. See Section 4. □

Control of the deviation of $\|A^*\|_2$ by the Matrix Hoeffding Inequality

We can write $\|A^*\|_2$ as

$$\|A^*\|_2 \quad = \quad \Big\| \sum_{j^* \in T^*} A^*_{j^*} \Big\|,$$

where $A^*_{j^*}$ is the matrix

$$A^*_{j^*} \quad = \quad W_{j^*} \begin{bmatrix} 0 & \mathfrak{C}^t_{k_{j^*}} \beta^*_{j^*} \\ \mathfrak{C}_{k_{j^*}} \beta^*_{j^*} & 0 \end{bmatrix}$$

Thus, by the triangular inequality, we have

$$\|A^*\|_2 \quad \leq \quad \Big\| \sum_{j^* \in T^*} A^*_{j^*} - \mathbb{E}\left[A^*_{j^*} \mid \mathcal{E}^*_\alpha\right] \Big\| + \Big\| \sum_{j^* \in T^*} \mathbb{E}\left[A^*_{j^*} \mid \mathcal{E}^*_\alpha\right] \Big\|, \tag{4.-125}$$

and we may apply the Matrix Hoeffding inequality again. We have that

$$\|A^*_{j^*}\| \quad = \quad |W_{j^*}||\beta^*_{j^*}|$$

and thus, on $\mathcal{E}^*_\alpha$,

$$\|A^*_{j^*} - \mathbb{E}\left[A^*_{j^*} \mid \mathcal{E}^*_\alpha\right]\| \quad \leq \quad 2 \frac{\mathfrak{s}\sqrt{n \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}{1 - \mathfrak{s}\sqrt{n \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}} |\beta^*_{j^*}|.$$

Under Assumption 2.7, we have

$$\mathfrak{s}\sqrt{n \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}} \leq 0.1,$$

this former inequality becomes

$$\|A^*_{j^*} - \mathbb{E}\left[A^*_{j^*} \mid \mathcal{E}^*_\alpha\right]\| \quad \leq \quad 3\mathfrak{s}\sqrt{n \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}} |\beta_j|.$$

Applying the Matrix Hoeffding inequality, we obtain

$$\begin{aligned} &\mathbb{P}\left(\| \sum_{j^* \in T^*} A^*_{j^*} - \mathbb{E}\left[A^*_{j^*} \mid \mathcal{E}^*_\alpha\right] \|_2 \geq u \mid \mathcal{E}^*_\alpha \right) \\ &\qquad \leq 2(n+1) \cdot \exp\left(- \frac{u^2}{8 \cdot 9\ \mathfrak{s}^2 \left(n \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}} \right) \|\beta^*_{T^*}\|_2^2} \right) \end{aligned} \tag{4.-129}$$

Let us now turn to the expectation term, i.e. the last term in (4.-125). We have

$$\begin{aligned} \Big\| \sum_{j^* \in T^*} \mathbb{E}\left[A^*_{j^*} \mid \mathcal{E}^*_\alpha\right] \Big\| \quad &= \quad \Big\| \sum_{j^* \in T^*} \mathbb{E}\left[W_{j^*} \mid \mathcal{E}^*_\alpha\right] \begin{bmatrix} 0 & \mathfrak{C}^t_{k_{j^*}} \beta_{j^*} \\ \mathfrak{C}_{k_{j^*}} \beta_{j^*} & 0 \end{bmatrix} \Big\| \\ &\leq \quad \Big\| \sum_{j^* \in T^*} 3\mathfrak{s}\sqrt{n \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}} \begin{bmatrix} 0 & \mathfrak{C}^t_{k_{j^*}} \beta_{j^*} \\ \mathfrak{C}_{k_{j^*}} \beta_{j^*} & 0 \end{bmatrix} \Big\|. \end{aligned}$$

This last inequality, when combined with (4.-129) and (4.-125), implies

$$\mathbb{P}\left(\|A^*\|_2 \geq 3\mathfrak{s}\sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\|\mathfrak{C}_{\mathcal{K}_{T^*}}\beta_{T^*}\|_2 + u \mid \mathcal{E}_\alpha^*\right)$$
$$\leq 2(n+1)\cdot\exp\left(-\frac{u^2}{8\cdot 9\ \mathfrak{s}^2\left(n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\right)\|\beta_{T^*}^*\|_2^2}\right),$$

from which we deduce, by the same trick as in Section 4, that

$$\mathbb{P}\left(\|A^*\|_2 \geq 3\mathfrak{s}\sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\|\mathfrak{C}_{\mathcal{K}_{T^*}}\beta_{T^*}\|_2 + u\right)$$
$$\leq 2(n+1)\cdot\exp\left(-\frac{u^2}{8\cdot 9\ \mathfrak{s}^2\left(n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\right)\|\beta_{T^*}^*\|_2^2}\right) + \frac{1}{p^\alpha}.$$

Let us now choose u such that

$$2(n+1)\cdot\exp\left(-\frac{u^2}{8\cdot 9\ \mathfrak{s}^2\left(n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\right)\|\beta_{T^*}^*\|_2^2}\right),$$

i.e.

$$u \quad = \quad 8\sqrt{2}\ \mathfrak{s}\sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\|\beta_{T^*}^*\|_2\sqrt{\alpha\log(p)+\log(2n+2))}.$$

Therefore, we obtain that

$$\mathbb{P}\Bigg(\|A^*\|_2 \geq 3\mathfrak{s}\sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\|\mathfrak{C}_{\mathcal{K}_{T^*}}\beta_{T^*}\|_2$$
$$+8\sqrt{2}\ \mathfrak{s}\sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\|\beta_{T^*}^*\|_2\sqrt{\alpha\log(p)+\log(2n+2))}\Bigg)$$
$$\leq \frac{2}{p^\alpha}.$$

By the (3.-17), and the definition of β^*, we have

$$\begin{aligned}\|\beta_{T^*}^*\|_2 &\leq \sqrt{\rho_{\mathfrak{C}}}\ \|\mathfrak{C}_{\mathcal{K}_{T^*}}\beta_{T^*}\|_2,\\ &= \sqrt{\rho_{\mathfrak{C}}}\ \|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2\end{aligned}$$

and therefore, we obtain

$$\mathbb{P}\Bigg(\|A^*\|_2 \geq 3\mathfrak{s}\sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}} \left(1+2\sqrt{2}\ \sqrt{\rho_{\mathfrak{C}}}\sqrt{\alpha\log(p)+\log(2n+2))}\right)\|\mathfrak{C}_{\mathcal{K}_T}\beta_T\|_2\Bigg) \leq \frac{2}{p^\alpha}.$$

Proof of Lemma 4.1

Using the independence of the E_j, $j \in \mathcal{J}_{k_{j^*}}$, we have

$$\begin{aligned}\mathbb{P}\left(\|E_{j^*}\|_2 \geq u\right) &= \mathbb{P}\left(\min_{j\in\mathcal{J}_{k_{j^*}}}\|E_j\|_2 \geq u\right)\\ &= \prod_{j\in\mathcal{J}_{k_{j^*}}}\mathbb{P}\left(\|E_j\|_2^2 \geq u^2\right),\\ &\leq \mathbb{P}\left(\|E_j\|_2^2 \geq u^2\right)^{\min_{j^*\in T^*}\ |\mathcal{J}_{k_{j^*}}|}.\end{aligned}$$

We also have

$$\mathbb{P}\left(\|E_j\|_2^2 \geq u^2\right) = 1-\mathbb{P}\left(\|E_j\|_2^2 \leq u^2\right).$$

On the other hand, as is well known, we have

$$\mathbb{P}\left(\frac{\|E_j\|_2^2}{\mathfrak{s}^2} \leq u^2\right) \leq C_\chi\left(\frac{u^2}{n}\right)^n$$

for some positive constant C_χ. Thus, the union bound gives

$$\mathbb{P}\left(\max_{j^*\in T^*}\ \|E_{j^*}\|_2 \geq u\right) \leq s^*\left(1-C_\chi\left(\frac{u^2}{n\ \mathfrak{s}^2}\right)^n\right)^{\min_{j^*\in T^*}\ |\mathcal{J}_{k_{j^*}}|}.$$

Let us tune u so that

$$s^*\left(1-C_\chi\left(\frac{u^2}{n\ \mathfrak{s}^2}\right)^n\right) \leq \frac{1}{p^\alpha}$$

i.e.

$$u^2 \geq \frac{n\ \mathfrak{s}^2}{C_\chi^{\frac{1}{n}}}\left(1-\left(s^*p^{-\alpha}\right)^{\frac{1}{\min_{j^*\in T^*}|\mathcal{J}_{k_{j^*}}|}}\right)^{\frac{1}{n}}$$

and since $\min_{j^*\in T^*}|\mathcal{J}_{k_{j^*}}| \geq \vartheta_*\log(p)^\nu$,

$$u^2 \geq \frac{n\ \mathfrak{s}^2}{C_\chi^{\frac{1}{n}}}\left(1-\exp\left(-\frac{\alpha}{\vartheta_*\log(p)^{\nu-1}}-\frac{\log(s^*)}{\vartheta_*\log(p)^\nu}\right)\right)^{\frac{1}{n}}.$$

On $(0,1)$, we have

$$\exp(-z) \leq 1-(1-e^{-1})z$$

and thus,

$$u^2 \quad \geq \quad n\,\mathfrak{s}^2\,\left(\frac{\alpha\,(1-e^{-1})}{\vartheta_*\,C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}},$$

from which the desired estimate follows.

Proof of Lemma 3.6

Concentration of $\|B^*\|_2$

We start with the concentration of

$$\|B^*\|_2 \quad = \quad \Big\|\sum_{j^*\in T^*}\frac{\beta^*_{j^*}}{\|\mathfrak{C}_{k_{j^*}}+E_{j^*}\|_2}E_{j^*}\Big\|_2.$$

Consider the matrix $E^t_{T^*}$, whose columns are independent. We would like to bound its operator norm.

Lemma 4.2 *We have*

$$\mathbb{P}\left(\left\|E^t_{T^*}\right\|\geq \mathfrak{s}K_{n,s^*}\mid \mathcal{E}^*_\alpha\right) \quad \leq \quad \frac{2}{p^\alpha}$$

where we recall that K_{n,s^*} is defined by (2.-6) above.

extbfProof. See Section 4. □ Define

$$\mathcal{F}^*_\alpha \quad = \quad \mathcal{E}^*_\alpha\cap\left\{\left\|E^t_{T^*}\right\|\leq \mathfrak{s}K_{n,s^*}\right\}.$$

Thus, the union bound gives that $\mathbb{P}\left(\mathcal{F}^*_\alpha\right)\geq 3/p^\alpha$. Let us now turn to the task of bounding $\|B^*\|_2$. Notice that on $\mathcal{F}^*_\alpha$, we have

$$\left\|\sum_{j^*\in T^*}\frac{\beta^*_{j^*}}{\|\mathfrak{C}_{k_{j^*}}+E_{j^*}\|_2}E_{j^*}\right\|_2 \quad \leq \quad \max_b\left\|\sum_{j^*\in T^*}\frac{\beta^*_{j^*}}{b}E_{j^*}\right\|_2,$$

where the maximum is over all

$$b\in\left[1-\mathfrak{s}\sqrt{n\,\left(\frac{\alpha\,(1-e^{-1})}{\vartheta_*\,C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}},1+\mathfrak{s}\sqrt{n\,\left(\frac{\alpha\,(1-e^{-1})}{\vartheta_*\,C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\right].$$

Thus, on $\mathcal{F}^*_\alpha$,

$$\left\|\sum_{j^*\in T^*}\frac{\beta^*_{j^*}}{\|\mathfrak{C}_{k_{j^*}}+E_{j^*}\|_2}E_{j^*}\right\|_2 \quad \leq \quad \max_{b,\|w\|_2=1}\langle w,\sum_{j^*\in T^*}\frac{\beta^*_{j^*}}{b}E_{i,j^*}\rangle.$$

Now, we have

$$\mathbb{P}\left(\|B^*\|_2-\mathbb{E}\left[\max_{b,\|w\|_2=1}\langle w,\sum_{j^*\in T^*}\frac{\beta^*_{j^*}}{b}E_{j^*}\rangle\mid\mathcal{F}^*_\alpha\right]\geq u\mid\mathcal{F}^*_\alpha\right)$$
$$\leq\mathbb{P}\left(\max_{b,\|w\|_2=1}\langle w,\sum_{j^*\in T^*}\frac{\beta^*_{j^*}}{b}E_{j^*}\rangle-\mathbb{E}\left[\max_{b,\|w\|_2=1}\langle w,\sum_{j^*\in T^*}\frac{\beta^*_{j^*}}{b}E_{j^*}\rangle\mid\mathcal{F}^*_\alpha\right]\geq u\mid\mathcal{F}^*_\alpha\right).$$

Let

$$M^*_{b,w} \quad = \quad \langle w, \sum_{j^* \in T^*} \frac{\beta^*_{j^*}}{b} E_{j^*} \rangle.$$

We again have to bound $M^*_{b,w}$ on $\mathcal{F}^*_\alpha$, and its conditional variance given $\mathcal{F}^*_\alpha$. The Cauchy-Schwartz inequality gives

$$M^*_{b,w} \quad \leq \quad \frac{1}{b} \|\beta^*_{T^*}\|_2 \sqrt{\sum_{j^* \in T^*} (w^t E_{j^*})^2},$$

and thus,

$$\begin{aligned} M^*_{b,w} \quad &\leq \quad \frac{1}{b} \|\beta^*_{T^*}\|_2 \|E^t_{T^*} w\|_2 \\ &\leq \quad \frac{1}{b} \|\beta^*_{T^*}\|_2 \|E^t_{T^*}\| \|w\|_2. \end{aligned}$$

Thus, on $\mathcal{F}^*_\alpha$, using the fact that $\|w\|_2 = 1$, we have

$$M^*_{b,w} \quad \leq \quad \mu^*_{\max} \|\beta^*_{T^*}\|_2,$$

where

$$\mu^*_{\max} \quad = \quad \mathfrak{s} K_{n,s^*},$$

and K_{n,s^*} is defined by (2.-6). Let us now turn to the conditional variance of $M^*_{b,w}$ given $\mathcal{F}^*_\alpha$.

Lemma 4.3 *We have*

$$Var\left(M^*_{b,w} \mid \mathcal{F}^*_\alpha\right) \quad \leq \quad {\sigma^*_{\max}}^2$$

where ${\sigma^*_{\max}}^2$ *is defined by (2.-6).*

extbfProof. See Section 4. □ Using Talagrand's inequality (Theorem 5.6) again, we obtain that

$$\begin{aligned} \mathbb{P}\left(\max_{b, \|w\|_2 = 1} \frac{M^*_{b,w}}{\mu^*_{\max} \|\beta^*_{T^*}\|_2} \geq \mathbb{E}\left[\max_{b, \|w\|_2 = 1} \frac{M^*_{b,w}}{\mu^*_{\max} \|\beta^*_{T^*}\|_2} \mid \mathcal{F}^*_\alpha \right] + \sqrt{2u\gamma^*} + \frac{u}{3} \mid \mathcal{F}^*_\alpha \right) \\ \leq \exp(-u), \end{aligned} \tag{4.A70}$$

with

$$\gamma^* = n \frac{{\sigma^*_{\max}}^2}{{\mu^*_{\max}}^2 \|\beta^*_{T^*}\|_2^2} + \mathbb{E}\left[\max_{b, \|w\|_2 = 1} \frac{M^*_{b,w}}{\mu^*_{\max} \|\beta^*_{T^*}\|_2} \mid \mathcal{F}^*_\alpha \right].$$

Control of the conditional expectation of $\max_{b, \|w\|_2 = 1} \frac{M^*_{b,w}}{\mu^*_{\max} \|\beta^*_{T^*}\|_2}$

As in Section 4, we will use Dudley's entropy integral bound to control the expectation, but this time, the sub-Gaussian version of Section 5.9. Let us rewrite

$$\mathbb{E}\left[\max_{b, \|w\|_2 = 1} M^*_{b,w} \mid \mathcal{F}^*_\alpha \right] \quad = \quad \mathbb{E}\left[\max_{\tilde{w} \in \mathcal{T}^*} \langle \tilde{w}, \sum_{j^* \in T^*} \beta^*_{j^*} E_{j^*} \rangle \mid \mathcal{F}^*_\alpha \right].$$

First, we have to prove the sub-Gaussianity of $M^*_{b,w}$. Notice that, due to rotational invariance of the Gaussian measure, conditionally on $\mathcal{F}^*_\alpha$, $E^t_{j^*} w$ is centered and

$$\begin{aligned}\mathbb{P}\left(\langle \tilde{w}-\tilde{w}', \sum_{j^*\in T^*} \beta^*_{j^*} E_{j^*}\rangle \geq u \mid \mathcal{F}^*_\alpha\right) &\leq \mathbb{P}\left(\sum_{j^*\in T^*} \beta^*_{j^*} \left(\tilde{w}-\tilde{w}'\right)^t E_{j^*} \geq u \mid \mathcal{F}^*_\alpha\right)\\ &= \mathbb{P}\left(\sum_{j^*\in T^*} \beta^*_{j^*} (O_{\tilde{w}-\tilde{w}'} D(\zeta) E_{j^*})^t (\tilde{w}-\tilde{w}') \geq u \mid \mathcal{F}^*_\alpha\right).\end{aligned}$$

where ζ is a rademacher ± 1 random vector, $O_{\tilde{w}-\tilde{w}'}$ is the orthogonal transform which sends $\tilde{w}-\tilde{w}'$ to the vector $\|\tilde{w}-\tilde{w}'\|_2/\sqrt{n} e$, where e is the vector of all ones. Thus,

$$\mathbb{P}\left(\langle \tilde{w}-\tilde{w}', \sum_{j^*\in T^*} \beta^*_{j^*} E_{j^*}\rangle \geq u \mid \mathcal{F}^*_\alpha\right) = \mathbb{P}\left(\frac{\|\tilde{w}-\tilde{w}'\|_2}{\sqrt{n}} \sum_{j^*\in T^*} \beta^*_{j^*} \sum_{i=1}^n \zeta_i E_{i,j^*} \geq u \mid \mathcal{F}^*_\alpha\right).$$

We now study the sub-Gaussianity of $\sum_{i=1}^n \zeta_i E_{i,j^*}$. Using the Laplace transform version of Hoeffding's inequality, we have

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\eta \frac{\|\tilde{w}-\tilde{w}'\|_2}{\sqrt{n}} \sum_{i=1}^n \zeta_i \ {\beta^*}_{j^*} \ E_{i,j^*}\right) \mid E_{j^*}, \mathcal{F}^*_\alpha\right]\\ &\qquad \leq \exp\left(\eta^2 \ \frac{\|\tilde{w}-\tilde{w}'\|_2^2}{n} {\beta^*}^2_{j^*} \ \mathfrak{s}^2 \left(n \ \left(\frac{\alpha \ (1-e^{-1})}{\vartheta_* \ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\right)\right).\end{aligned}$$

Therefore, using independence of the E_{j^*}'s, we have that

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\frac{\eta \ \|\tilde{w}-\tilde{w}'\|_2}{\sqrt{n}} \sum_{j^*\in \mathcal{T}^*} {\beta^*}_{j^*} \sum_{i=1}^n \zeta_i E_{i,j^*}\right) \mid \mathcal{F}^*_\alpha\right]\\ &\qquad = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\frac{\eta \ \|\tilde{w}-\tilde{w}'\|_2}{\sqrt{n}} \sum_{j^*\in \mathcal{T}^*} {\beta^*}_{j^*} \sum_{i=1}^n \zeta_i E_{i,j^*}\right) \mid E_{j^*}, \mathcal{F}^*_\alpha\right] \mid \mathcal{F}^*_\alpha\right]\\ &\qquad = \mathbb{E}\left[\prod_{j^*\in T^*} \mathbb{E}\left[\exp\left(\frac{\eta \ \|\tilde{w}-\tilde{w}'\|_2}{\sqrt{n}} {\beta^*}_{j^*} \sum_{i=1}^n \zeta_i E_{i,j^*}\right) \mid E_{j^*}, \mathcal{F}^*_\alpha\right] \mid \mathcal{F}^*_\alpha\right]\\ &\qquad \leq \exp\left(\eta^2 \ \|\tilde{w}-\tilde{w}'\|_2^2 \|{\beta^*}_{T^*}\|_2^2 \ \mathfrak{s}^2 \left(\frac{\alpha \ (1-e^{-1})}{\vartheta_* \ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\right).\end{aligned}$$

Now Chernov's bound gives

$$\begin{aligned}&\mathbb{P}\left(\frac{\|\tilde{w}-\tilde{w}'\|_2}{\sqrt{n}} \sum_{j^*\in \mathcal{T}^*} {\beta^*}_{j^*} \sum_{i=1}^n \zeta_i E_{i,j^*} \geq u \mid \mathcal{F}^*_\alpha\right)\\ &\qquad \leq e^{-\eta u} \ \mathbb{E}\left[\exp\left(\frac{\eta \ \|\tilde{w}-\tilde{w}'\|_2}{\sqrt{n}} \sum_{j^*\in \mathcal{T}^*} {\beta^*}_{j^*} \sum_{i=1}^n \zeta_i E_{i,j^*}\right) \mid \mathcal{F}^*_\alpha\right]\\ &\qquad \leq \exp\left(\eta^2 \ \|\tilde{w}-\tilde{w}'\|_2^2 \|{\beta^*}_{T^*}\|_2^2 \ \mathfrak{s}^2 \left(\frac{\alpha \ (1-e^{-1})}{\vartheta_* \ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}} - \eta u\right).\end{aligned}$$

Optimizing in η gives

$$\mathbb{P}\left(\frac{\|\tilde{w}-\tilde{w}'\|_2}{\sqrt{n}}\sum_{j^*\in\mathcal{T}^*}{\beta^*}_{j^*}\sum_{i=1}^{n}\zeta_i E_{i,j^*}\geq u\mid\mathcal{F}_\alpha^*\right)$$
$$\leq\exp\left(-\frac{1}{4}\frac{u^2}{\|\tilde{w}-\tilde{w}'\|_2^2\ \|{\beta^*}_{T^*}\|_2^2\ \mathfrak{s}^2\left(\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\right)}\right).$$

Using the union bound and invariance of the bound with respect to sign change, we thus obtain

$$\mathbb{P}\left(\left|\langle\tilde{w}-\tilde{w}',\sum_{j^*\in T^*}\beta_{j^*}^*E_{j^*}\rangle\right|\geq u\mid\mathcal{F}_\alpha^*\right)$$
$$\leq 2\ \exp\left(-\frac{1}{4}\frac{u^2}{\|\tilde{w}-\tilde{w}'\|_2^2\ \|{\beta^*}_{T^*}\|_2^2\ \mathfrak{s}^2\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\right).$$

Thus, the process is sub-Gaussian with the semi-metric d, given by

$$d^2(\tilde{w},\tilde{w}') \quad=\quad 4\|\tilde{w}-\tilde{w}'\|_2^2\ \|{\beta^*}_{T^*}\|_2^2\ \mathfrak{s}^2\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}.$$

Let us now apply Theorem 5.9. The diameter of $\mathcal{T}^*$ is bounded from above by

$$r_{\max}^* \quad=\quad \frac{1}{1-\mathfrak{s}\sqrt{n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}.$$

An upper bound on the covering number of $\mathcal{T}^*$ with respect to the semi-metric d is given by

$$H(\varepsilon,\mathcal{T}^*) \quad\leq\quad n\log\left(\frac{3\cdot 2r_{\max}^*\ \|{\beta^*}_{T^*}\|_2\ \mathfrak{s}\sqrt{\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}{\varepsilon}\right).$$

Therefore, by Theorem 5.7, we obtain that

$$\mathbb{E}\left[\max_{\tilde{w}\in\mathcal{T}^*}\langle\tilde{w},\sum_{j\in T}\beta_j E_j\rangle\mid\mathcal{F}_\alpha^*\right] \quad\leq\quad 12\sqrt{n}\int_0^{\sigma_G}\sqrt{\log\left(\frac{6\ r_{\max}^*\ \|{\beta^*}_{T^*}\|_2\ \mathfrak{s}\sqrt{\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}{\varepsilon}\right)}d\varepsilon,$$

with

$$\sigma_G \quad=\quad r_{\max}^*.$$

Using the change of variable

$$\varepsilon'=\frac{\varepsilon}{2\ r_{\max}^*\ \|{\beta^*}_{T^*}\|_2\ \mathfrak{s}\sqrt{\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}},$$

we obtain

$$\mathbb{E}\left[\max_{\tilde{w}\in\mathcal{T}}\langle\tilde{w},\sum_{j\in T}\beta_j E_j\rangle \mid \mathcal{F}_\alpha^*\right] \leq 24\, r_{\max}^*\, \|\beta^*{}_{T^*}\|_2\, \mathfrak{s}\sqrt{\left(\frac{\alpha\,(1-e^{-1})}{\vartheta_*\, C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\, C_\int^* \quad (4.\text{-}196)$$

where

$$C_\int^* \;=\; \int_0^3 \sqrt{\log\left(\frac{3}{\varepsilon'}\right)}\, d\varepsilon'.$$

Last step of the proof

Combining (4.-170) and (4.-196), we obtain

$$\mathbb{P}\Bigg(\|B^*\| \geq 24\, r_{\max}^*\, \|\beta^*{}_{T^*}\|_2\, \mathfrak{s}\sqrt{\left(\frac{\alpha\,(1-e^{-1})}{\vartheta_*\, C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\, C_\int^* + \mu_{\max}^*\, \|\beta_{T^*}^*\|_2\left(\sqrt{2u\gamma^*}+\frac{u}{3}\right) \mid \mathcal{F}_\alpha^*\Bigg) \leq \exp(-u) \quad (4.\text{-}197)$$

with

$$\gamma^* = n\frac{{\sigma_{\max}^*}^2}{{\mu_{\max}^*}^2\, \|\beta_{T^*}^*\|_2^2} + \mathbb{E}\left[\max_{b,\|w\|_2=1}\frac{M_{b,w}^*}{\mu_{\max}^*\, \|\beta_{T^*}^*\|_2} \mid \mathcal{F}_\alpha^*\right].$$

Therefore, taking $u = \alpha\log(p)$ we have

$$\mathbb{P}\Bigg(\|B^*\| \geq 24\, r_{\max}^*\, \|\beta^*{}_{T^*}\|_2\, \mathfrak{s}\sqrt{\left(\frac{\alpha\,(1-e^{-1})}{\vartheta_*\, C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\, C_\int^* + \left(\sqrt{2\alpha\log(p)\,(L^*)}+\frac{\alpha\log(p)}{3}\mu_{\max}^*\right)\|\beta_{T^*}^*\|_2 \mid \mathcal{F}_\alpha^*\Bigg) \leq p^{-\alpha},$$

with

$$L^* \;=\; n\frac{{\sigma_{\max}^*}^2}{\|\beta^*{}_{T^*}\|_2^2} + 24\,\mu_{\max}^* r_{\max}^*\, \mathfrak{s}\sqrt{\left(\frac{\alpha\,(1-e^{-1})}{\vartheta_*\, C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\, C_\int^*.$$

Using Assumption 2.9, we obtain

$$\mathbb{P}\Bigg(\|B^*\| \geq \left(24\, r_{\max}^*\, \mathfrak{s}\sqrt{\left(\frac{\alpha\,(1-e^{-1})}{\vartheta_*\, C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\, C_\int^* + \mu_{\max}^*\,\alpha\log(p)\right)\|\beta_{T^*}^*\|_2 \mid \mathcal{F}_\alpha^*\Bigg) \leq p^{-\alpha}.$$

Using the same trick as before, we obtain

$$\mathbb{P}\left(\|B^*\| \geq \left(24\ r^*_{\max}\ \mathfrak{s}\sqrt{\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\ C^*_\int \right.\right.$$
$$\left.\left. +\mu^*_{\max}\ \alpha\log(p)\right)\ \|\beta^*_{T^*}\|_2\right) \leq \frac{2}{p^\alpha}.$$

Notice further that

$$\begin{aligned}\|\beta^*_{T^*}\|_2 &\leq (1+\rho_{\mathfrak{C}})\|\mathfrak{C}_{\mathcal{K}_{T^*}}\beta^*_{T^*}\|_2\\ &= (1+\rho_{\mathfrak{C}})\|\mathfrak{C}_T\beta_T\|_2,\end{aligned}$$

by definition of β^*. Thus, we obtain that

$$\mathbb{P}\left(\|B^*\| \geq \left(24\ r^*_{\max}\ \mathfrak{s}\sqrt{\left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\ C^*_\int \right.\right.$$
$$\left.\left. +\mu^*_{\max}\ \alpha\log(p)\right)\ \sqrt{\rho_{\mathfrak{C}}}\ \|\mathfrak{C}_T\beta_T\|_2\right) \leq \frac{2}{p^\alpha}.$$

as desired.

Proof of Lemma 4.2

Let us first notice that since $\|E_{T^*}\| = \|E^t_{T^*}\|$, we can write

$$\begin{aligned}\|E^t_{T^*}\| &= \sqrt{\|E_{T^*}E^t_{T^*}\|}\\ &= \sqrt{\left\|\sum_{j^*\in T^*} E_{j^*}E^t_{j^*}\right\|}\end{aligned}$$

This latter expression is well suited for our problem, since it is the norm of the sum of independent positive semi-definite random matrices, for which the Matrix Chernov inequality of Section 5 applies. In order to apply this inequality, we need a bound on the norm of each summand. By Lemma 4.1, on $\mathcal{E}^*$, we have

$$\begin{aligned}\left\|E_{j^*}E^t_{j^*}\right\| &= \|E_{j^*}\|_2^2\\ &\leq \mathfrak{s}^2\ n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}.\end{aligned}$$

We also need a bound on the norm of the expectation. We have

$$\left\|\mathbb{E}\left[\sum_{j^*\in T^*} E_{j^*}E^t_{j^*}\mid \mathcal{F}^*_\alpha\right]\right\| = \left\|\sum_{j^*\in T^*}\mathbb{E}\left[E_{j^*}E^t_{j^*}\mid \mathcal{F}^*_\alpha\right]\right\|.$$

Due to rotational invariance, we have that the law of E_{j^*} is the same as the law of $D(\zeta)E_{j^*}$, where $\zeta_1,\ldots,\zeta_n$ are i.i.d. Rademacher ±1 random variables independent from E_{j^*}. Thus,

$$\begin{aligned}\mathbb{E}\left[\zeta_i E_{i,j^*}\zeta_{i'}E_{i',j^*}\mid \mathcal{E}^*_\alpha\right] &= \mathbb{E}\left[\mathbb{E}\left[\zeta_i E_{i,j^*}\zeta_{i'}E_{i',j^*}\mid E_{i,j^*},\ E_{i',j^*}\mid \mathcal{E}^*_\alpha\right]\right]\\ &= 0. \qquad (4.\text{-}215)\end{aligned}$$

On the other hand, we have the following result.

Lemma 4.4 *We have*

$$\mathbb{E}\left[E_{i,j^*}^2 \mid \mathcal{E}_\alpha^*\right] \leq \mathfrak{s}^2 \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}.$$

extbfProof. Due to rotational invariance of the law of E_{j^*} and the event $\mathcal{E}_\alpha^*$, we have

$$\mathbb{E}\left[E_{1,j^*}^2 \mid \mathcal{E}_\alpha^*\right] = \cdots = \mathbb{E}\left[E_{n,j^*}^2 \mid \mathcal{E}_\alpha^*\right].$$

Therefore,

$$\mathbb{E}\left[E_{i,j^*}^2 \mid \mathcal{E}_\alpha^*\right] \leq \frac{1}{n}\,\mathbb{E}\left[\sum_{i'=1}^{n} E_{i',j^*}^2 \mid \mathcal{E}_\alpha^*\right]$$

and by the definition of $\mathcal{E}_\alpha^*$,

$$\mathbb{E}\left[E_{i,j^*}^2 \mid \mathcal{E}_\alpha^*\right] = \mathfrak{s}^2 \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}.$$

□

Based on this lemma, and the fact that the matrix

$$\mathbb{E}\left[\sum_{j^*\in T^*} E_{j^*}E_{j^*}^t \mid \mathcal{F}_\alpha^*\right],$$

is diagonal by (4.-215), we obviously obtain that

$$\left\|\mathbb{E}\left[\sum_{j^*\in T^*} E_{j^*}E_{j^*}^t \mid \mathcal{F}_\alpha^*\right]\right\| = \mathfrak{s}^2 \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}.$$

With the bound on the norm of the expectation and on the variance in hand, we are now ready to apply the Matrix Chernov inequality and obtain

$$\mathbb{P}\left(\left\|\sum_{j^*\in T^*} E_{j^*}E_{j^*}^t\right\| \geq u \mid \mathcal{F}_\alpha^*\right) \leq n\left(\frac{e\ \mathfrak{s}^2\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}{u}\right)^{\frac{u}{\mathfrak{s}^2\ n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}.$$

Let us finally tune u so that the right hand side term is less than $p^{-\alpha}$, i.e.

$$\log(n) + \log\left(\frac{e\ \mathfrak{s}^2\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}{u}\right) \leq -\alpha\ \frac{\mathfrak{s}^2\ n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}{u}\ \log(p).$$

Take

$$u \quad = \quad \alpha\ \mathfrak{s}^2\ n\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}} \left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}\ \log(p). \tag{4.-224}$$

Since, by assumption, $p \geq e^{e^{2-\log(\alpha)}}$, we have $-\log(\log(p)) + \log(e/\alpha) \leq -1$. Moreover, the value of u given by (4.-224) is less than or equal to $\mathfrak{s}^2\ K^2_{n,s^*}$ with K_{n,s^*} given by (2.-6). This completes the proof.

Proof of Lemma 4.3

Independence of the E_{j^*}, $j^* \in T^*$ allows to write

$$Var\left(M^*_{b,w} \mid \mathcal{F}^*_\alpha\right) \quad = \quad \sum_{j^* \in T^*} \frac{\beta^*_{j^*}}{b} Var\left(E^t_{j^*} w \mid \mathcal{F}^*_\alpha\right),$$

and, using the Cauchy-Schwartz inequality again, we obtain

$$Var\left(M^*_{b,w} \mid \mathcal{F}^*_\alpha\right) \quad = \quad \frac{\|\beta^*_{T^*}\|_2}{b} \sqrt{\sum_{j^* \in T^*} Var^2\left(E^t_{j^*} w \mid \mathcal{F}^*_\alpha\right)}.$$

On the other hand, notice that, due to rotational invariance of the Gaussian measure, conditionally on $\mathcal{F}^*_\alpha$, $E^t_{j^*} w$ is centered and

$$\begin{aligned} Var\left(E^t_{j^*} w \mid \mathcal{F}^*_\alpha\right) \quad &= \quad \mathbb{E}\left[\left((O_w D(\zeta) E_{j^*})^t w\right)^2 \mid \mathcal{F}^*_\alpha\right], \\ &= \quad \mathbb{E}\left[\left(E^t_{j^*} D(\zeta) O_w w\right)^2 \mid \mathcal{F}^*_\alpha\right], \end{aligned}$$

where ζ is a rademacher ± 1 random vector, O_w is the orthogonal transform which sends w to the vector $1/\sqrt{n}e$, where e is the vector of all ones. Thus,

$$Var\left(E^t_{j^*} w \mid \mathcal{F}^*_\alpha\right) \quad = \quad \frac{1}{n}\mathbb{E}\left[\mathbb{E}\left[\left(E^t_{j^*} D(\zeta) e\right)^2 \mid E, \mathcal{F}^*_\alpha\right] \mid \mathcal{F}^*_\alpha\right],$$

Moreover,

$$\mathbb{E}\left[\left(E^t_{j^*} D(\zeta) e\right)^2 \mid E, \mathcal{F}^*_\alpha\right] \quad = \quad \mathbb{E}\left[\left(\sum_{i=1}^n E_{i,j^*} \zeta_i\right)^2 \mid E, \mathcal{F}^*_\alpha\right]$$

and expanding the square of the sum gives

$$\mathbb{E}\left[\left(E^t_{j^*} D(\zeta) e\right)^2 \mid E, \mathcal{F}^*_\alpha\right] \quad = \quad \|E_{j^*}\|_2^2.$$

Using the bound on b, we finally obtain

$$Var\left(M^*_{b,w} \mid \mathcal{F}^*_\alpha\right) \quad \leq \quad {\sigma^*_{\max}}^2,$$

where ${\sigma^*_{\max}}^2$ is given by (2.-6).

5 Norms of random matrices, ε-nets and concentration inequalities

Norms and coverings

Proposition 5.1 *([22, Proposition 2.1]). For any positive integer d, there exists an ε-net of the unit sphere of $\mathbb{R}^d$ of cardinality*

$$2d\left(1+\frac{2}{\varepsilon}\right)^{d-1} \quad \leq \quad \left(\frac{3}{\varepsilon}\right)^{d}.$$

The next proposition controls the approximation of the norm based on an ε-net.

Proposition 5.2 *([22, Proposition 2.2]). Let $\mathcal{N}$ be an ε-net of the unit sphere of $\mathbb{R}^d$ and let $\mathcal{N}'$ be an ε'-net of the unit sphere of $\mathbb{R}^{d'}$. Then for any linear operator $A : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$, we have*

$$\|A\| \quad \leq \quad \frac{1}{(1-\varepsilon)(1-\varepsilon')} \sup_{\substack{v\in\mathcal{N}\\ w\in\mathcal{N}'}} |v^t A w|.$$

The Matrix Hoeffding Inequality

A Non-commutative version of the famous Hoeffding inequality was proposed in [25]. We recall this result for convenience.

Theorem 5.3 *Consider a finite sequence $(U_j)_{j\in T}$ of independent random, self-adjoint matrices with dimension d, and let $(U_j)_{j\in T}$ be a sequence of deterministic self-adjoint matrices. Assume that each random matrix satisfies*

$$\mathbb{E}\,[U_j] \quad = 0 \quad \textit{and} \quad U_j^2 \preceq \quad V_j^2 \ \textit{a.s.}$$

for all $j \in T$. Then, for all $u \geq 0$,

$$\mathbb{P}\left(\lambda_{\max}\Big(\sum_{j\in T} U_j\Big) \geq t\right) \quad \leq \quad d\cdot\exp\left(-\frac{u^2}{8\left\|\sum_{j\in T} V_j^2\right\|}\right).$$

The Matrix Chernov inequality

The following non-commutative version of Chernoff's inequality was recently established in [25].

Theorem 5.4 *(Matrix Chernoff Inequality [25]) Let $X_1,\ldots,X_p$ be independent random positive semi-definite matrices taking values in $\mathbb{R}^{d\times d}$. Set $S_p = \sum_{j=1}^p X_j$. Assume that for all $j \in \{1,\ldots,p\}$ $\|X_j\| \leq B$ a.s. and*

$$\|\mathbb{E}\ S_p\| \quad \leq \quad \mu_{\max}.$$

Then, for all $r \geq e\ \mu_{\max}$,

$$\mathbb{P}\left(\|S_p\| \geq r\right) \quad \leq \quad d\left(\frac{e\ \mu_{\max}}{r}\right)^{r/B}.$$

(Set $r = (1+\delta)\mu_{\max}$ and use $e^\delta \leq e^{1+\delta}$ in Theorem 1.1 [25].)

Gaussian i.i.d. matrices

The following result on random matrices with Gaussian i.i.d. entries can be found in [27, Corollary 5.35].

Theorem 5.5 *Let G be an $n \times m$ be a matrix whose entries are independent standard normal random variables. Then for every $u \geq 0$, with probability at least $1 - 2\exp\left(-u^2/2\right)$, one has*

$$\sqrt{n} - \sqrt{m} - u \leq \quad \sigma_{\min}(G) \leq \sigma_{\max}(G) \quad \leq \sqrt{n} + \sqrt{m} + u.$$

Talagrand's concentration inequality for empirical processes

The following theorem, which a version of Talagrand's concentration inequality for empirical processes, was proved in [4, Theorem 2.3].

Theorem 5.6 *Let X_i be a sequence of i.i.d. variables taking values in a Polish space $\mathcal{X}$, and let $\mathcal{F}$ be a countable family of functions from $\mathcal{X}$ to $\mathbb{R}$ and assume that all functions f in $\mathcal{F}$ are measurable, square integrable and satisfy $\mathbb{E}[f] = 0$. If $\sup_{f\in\mathcal{F}} ess\sup f \leq 1$, then we denote*

$$Z \quad = \quad \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(X_i).$$

Let $\sigma_{\max}$ be a positive number such that $\sigma_{\max}^2 \geq \sup_{f\in\mathcal{F}} Var\left(f(X_1)\right)$ almost surely, then, for all $u \geq 0$, we have

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + \sqrt{2u\gamma} + \frac{u}{3}\right) \quad \leq \quad \exp\left(-u\right),$$

with $\gamma = n\sigma_{\max}^2 + \mathbb{E}[Z]$.

Dudley's entropy integral bound

Let $(\mathcal{T}, d)$ denote a semi-metric space and denote by $H(\delta, \mathcal{T})$ the δ-entropy number of $(\mathcal{T}, d)$ for all positive real number δ.

The Gaussian case

Let $(G_t)_{t\in\mathcal{T}}$ be a centered gaussian process indexed by $\mathcal{T}$ and set d to be the covariance pseudo-metric defined by

$$d(t, t') \quad = \quad \sqrt{\mathbb{E}\left[(G_t - G_{t'})^2\right]}.$$

Then, we have the following important theorem of Dudley, which can be found in the present form in [18].

Theorem 5.7 *Assume that $(\mathcal{T}, d)$ is totally bounded. If $\sqrt{H(\delta, \mathcal{T})}$ is integrable at zero, then*

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}} G_t\right] \quad \leq \quad 12 \int_0^{\sigma_G} \sqrt{H(\delta, \mathcal{T})}\, d\delta,$$

where

$$\sigma_G^2 \quad = \quad \sup_{t\in\mathcal{T}} \mathbb{E}\left[G_t^2\right].$$

The sub-Gaussian case

We start with the definition of sub-Gaussian processes.

Definition 5.8 *A centered process $(S_t)_{t\in\mathcal{T}}$ is said to be sub-Gaussian if for all $(t,t')\in\mathcal{T}^2$, and for all $u>0$,*

$$\mathbb{P}\left(|X_t - X_{t'}| \geq u\right) \quad \leq \quad 2\exp\left(-\frac{u^2}{d^2(t,t')}\right).$$

One easily checks that a Gaussian process is sub-Gaussian with the covariance semi-metric in the former definition. Let $(S_t)_{t\in\mathcal{T}}$ be a centered sub-Gaussian process. We then have the following standard result.

Theorem 5.9 *Assume that $(\mathcal{T},d)$ is totally bounded. If $\sqrt{H(\delta,\mathcal{T})}$ is integrable at zero, then*

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}} S_t\right] \quad = \quad C_{chain}\int_0^{diam(\mathcal{T})}\sqrt{H(\delta,\mathcal{T})}\,d\delta$$

for some positive constant C_{chain}.

6 Verifying the Candes-Plan conditions

The goal of this section is to Proposition 3.1 which gives a version of Candès and Plan's conditions adapted to our Gaussian mixture model.

Important properties of $\mathfrak{C}$

The invertibily condition for (3.-17) is a direct consequence of [37]. An alternative approach, based on the Matrix Chernov inequality is proposed in [13], with improved constants. We have in particular

Theorem 6.1 *[13, Theorem 1] Let $r\in(0,1)$, $\alpha\geq 1$. Let Assumptions 2.2 and 2.4 hold with*

$$C_{spar} \quad \geq \quad \frac{r^2}{4(1+\alpha)e^2}. \tag{6.-245}$$

With $\mathcal{K}\subset\{1,\dots,K\}$ chosen randomly from the uniform distribution among subsets with cardinality s^*, the following bound holds:

$$\mathbb{P}\left(\|\mathfrak{C}_{\mathcal{K}}^t\mathfrak{C}_{\mathcal{K}} - I_s\| \geq r\right) \quad \leq \quad \frac{216}{p^\alpha}. \tag{6.-244}$$

Moreover, the following property will also be very useful.

Lemma 6.2 *(Adapted from [13, Lemma 5.3]) If $v^2 \geq e\ s^*\ \|\mathfrak{C}\|/K_o$, we have*

$$\mathbb{P}\left(\max_{k\in\mathcal{K}^c}\left\|\mathfrak{C}_{\mathcal{K}}^t\mathfrak{C}_k\right\| \geq \frac{v}{1-r}\right) \quad \leq \quad K_o\left(e\frac{s^*\ \|\mathfrak{C}\|^2}{K_o\ v^2}\right)^{\frac{v^2}{\mu(\mathfrak{C})^2}}.$$

Based on this lemma, we easily get the following bound.

Lemma 6.3 *Take* $C_{col} \geq e^2(\alpha+1)\max\{\sqrt{C_{spar}}, C_\mu\}/(1-r)$. *Then, we have*

$$\mathbb{P}\left(\max_{k\in\mathcal{K}^c}\left\|\mathfrak{C}_{\mathcal{K}}^t\mathfrak{C}_k\right\| \geq \frac{C_{col}}{(1-r)\sqrt{\log(p)}}\right) \quad \leq \quad \frac{1}{p^\alpha}.$$

extbfProof. Taking $v = C_{col}/\sqrt{\log(p)}$, we obtain from Lemma 6.2

$$\mathbb{P}\left(\max_{k\in\mathcal{K}^c}\left\|\mathfrak{C}_{\mathcal{K}}^t\mathfrak{C}_k\right\| \geq \frac{C_{col}}{(1-r)\sqrt{\log(p)}}\right) \quad \leq \quad K_o\left(e\frac{s^*\,\|\mathfrak{C}\|^2\log(p)}{K_o\,C_{col}^2}\right)^{\frac{C_{col}^2}{C_\mu^2}\log(p)}.$$

Using (2.4), this gives

$$\mathbb{P}\left(\max_{k\in\mathcal{K}^c}\left\|\mathfrak{C}_{\mathcal{K}}^t\mathfrak{C}_k\right\| \geq \frac{C_{col}}{(1-r)\sqrt{\log(p)}}\right) \quad \leq \quad K_o\left(e\frac{C_{spar}}{C_{col}^2}\right)^{\frac{C_{col}^2}{C_\mu^2}\log(p)}.$$

Since $C_{col} \geq e^2(\alpha+1)\max\{\sqrt{C_{spar}}, C_\mu\}$, we get

$$K_o\left(e\frac{C_{spar}}{C_{col}^2}\right)^{\frac{C_{col}^2}{C_\mu^2}\log(p)} \quad \leq \quad K_o\left(\frac{e}{\alpha+1}\right)^{(\alpha+1)\ \log(p)}$$

and since, by Assumption 2.1, $K_o \leq p$, we obtain that

$$\mathbb{P}\left(\max_{k\in\mathcal{K}^c}\left\|\mathfrak{C}_{\mathcal{K}}^t\mathfrak{C}_k\right\| \geq \frac{C_{col}}{(1-r)\sqrt{\log(p)}}\right) \quad \leq \quad \frac{1}{p^\alpha}.$$

□

Similar properties for X_{T^*}

Control of $\|X_{T^*}^t X_{T^*} - I\|$

We have

$$\sigma_{\min}\left(X_{T^*}^t X_{T^*}\right) \quad = \quad \sigma_{\min}\left(\left(\mathfrak{C}_{\mathcal{K}_{T^*}} + E_{T^*}\right)^t D_*^2\left(\mathfrak{C}_{\mathcal{K}_{T^*}} + E_{T^*}\right)\right)$$

where (see Step 1 in the proof of Proposition 3.2) D_* is a diagonal matrix whose diagonal elements are indexed by T^* and are defined by

$$D_{*,j^*,j^*} \quad = \quad \frac{1}{\left\|\mathfrak{C}_{k_{j^*}} + E_{j^*}\right\|_2},$$

for $j^* \in T^*$. By the definition of $\mathcal{E}_\alpha^*$, we have

$$\sigma_{\min}(D_*) \quad \geq \quad \frac{1}{1+\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}.$$

and

$$\sigma_{\max}(D_*) \quad \leq \quad \frac{1}{1-\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}.$$

By the triangular inequality,

$$\begin{aligned}\sigma_{\min}\left(X_{T^*}^t X_{T^*}\right) &\geq \sigma_{\min}\left(\mathfrak{C}_{\mathcal{K}}^t D_*^2 \mathfrak{C}_{\mathcal{K}}\right) - \left\|\mathfrak{C}_{\mathcal{K}}^t D_*^2 E_{T^*}\right\| - \left\|E_{T^*}^t D_*^2 E_{T^*}\right\| \\ &\geq \frac{1-r}{\left(1+\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\right)^2} \\ &\quad - \frac{(1+r)\ \|E_{T^*}\| + \|E_{T^*}\|^2}{\left(1-\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\right)^2}.\end{aligned}$$

and

$$\begin{aligned}\sigma_{\max}\left(X_{T^*}^t X_{T^*}\right) &\leq \left\|\mathfrak{C}_{\mathcal{K}}^t D_*^2 \mathfrak{C}_{\mathcal{K}}\right\| + \left\|\mathfrak{C}_{\mathcal{K}}^t D_*^2 E_{T^*}\right\| + \left\|E_{T^*}^t D_*^2 E_{T^*}\right\| \\ &\leq \frac{(1+r)+(1+r)\ \|E_{T^*}\| + \|E_{T^*}\|^2}{\left(1-\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\right)^2}.\end{aligned}$$

Moreover, using Theorem 6.1 and Lemma 4.2, we obtain

$$\mathbb{P}\left(\left\|X_{T^*}^t X_{T^*} - I\right\| \geq r^* \mid \mathcal{E}_\alpha^*\right) \leq \frac{218}{p^\alpha}$$

with r^* given by

$$\begin{aligned}r^* = \max\Bigg\{ &\frac{(1+r)+(1+r)\ \mathfrak{s}K_{n,s^*} + \mathfrak{s}^2 K_{n,s^*}^2}{\left(1-\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\right)^2} - 1; \\ &1 - \Bigg(\frac{1-r}{\left(1+\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\right)^2} \\ &- \frac{(1+r)\ \mathfrak{s}K_{n,s^*} + \mathfrak{s}^2 K_{n,s^*}^2}{\left(1-\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\right)^2}\Bigg)\Bigg\}.\end{aligned} \tag{6.-260}$$

Using Assumption (2.7), we have

$$\mathfrak{s}\ K_{n,s^*} \leq C_{\mathfrak{s},n,p}\frac{\sqrt{\alpha\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}{\left(1+\sqrt{\frac{\alpha+1}{c}\frac{\log(p)}{n}}\right)},$$

and thus, by Assumption 2.7,

$$\begin{aligned}\mathfrak{s}\ K_{n,s^*} &\leq C_{\mathfrak{s},n,p}\sqrt{\alpha\ \left(\frac{\alpha\ (1-e^{-1})}{\vartheta_*\ C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}, \\ &\leq 0.1\cdot r.\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\, C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}} &\leq \frac{C_{\mathfrak{s},n,p}}{\sqrt{\log(p)}}\,\frac{\sqrt{\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\, C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}}{\left(1+\sqrt{\frac{\frac{\alpha+1}{c}\log(p)}{n}}\right)}\\ &\leq \frac{C_{\mathfrak{s},n,p}}{\sqrt{\log(p)}}\,\sqrt{\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\, C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}}\end{aligned}$$

which, by Assumption 2.7, gives

$$\mathfrak{s}\sqrt{n\left(\frac{\alpha(1-e^{-1})}{\vartheta_*\, C_\chi}\right)^{\frac{1}{n}}\left(\frac{1}{\log(p)^{\nu-1}}\right)^{\frac{1}{n}}} \leq \frac{0.1\cdot r}{\log(p)}.$$

Summing up, we get

$$\begin{aligned}r^* &\leq \frac{(1+r)+(1+r)\;0.1\cdot r+0.01\cdot r^2}{\left(1-\frac{0.1\cdot r}{\log(p)}\right)^2}-1\\ &\leq \left(1.1\cdot r+0.11\cdot r^2\right)\\ &\qquad +2\;\left(1+1.1\cdot r+0.11\cdot r^2\right)\frac{0.1\cdot r}{\log(p)}\end{aligned}$$

and by Assumption 2.-8,

$$r^* \leq 1.1\cdot r\;\left(1.1+0.11\cdot r\right).$$

Thus, using Lemma 4.1,

$$\mathbb{P}\left(\left\|X_{T^*}^t X_{T^*}-I\right\|\geq 1.1\cdot r(1.1+0.11\cdot r)\right) \leq \frac{218+1}{p^\alpha},$$

Control of $\max_{k\in T^{*c}}\left\|X_{T^*}^t X_k\right\|_2$

By the triangular inequality, we have that

$$\begin{aligned}\max_{k\in T^{*c}}\left\|X_{T^*}^t X_k\right\|_2 &= \max_{k\in T^{*c}}\left\|\left(\mathfrak{C}_{\mathcal{K}}+E_{T^*}\right)^t D_*^2\left(\mathfrak{C}_k+E_k\right)\right\|_2\\ &\leq \left(\max_{k\in T^{*c}}\left\|\mathfrak{C}_{\mathcal{K}}^t\mathfrak{C}_k\right\|+\left\|\mathfrak{C}_{\mathcal{K}}\right\|\max_{k\in T^{*c}}\left\|E_k\right\|_2\right.\\ &\qquad\left.+\left\|E_{T^*}\right\|\max_{k\in T^{*c}}\left\|E_k\right\|_2\right)\left\|D_*\right\|^2.\end{aligned}$$

A computation analogous to the one for the probability of $\mathcal{E}_\alpha$ gives that

$$\mathbb{P}\left(\max_{k\in\{1,\ldots,p\}}\left\|E_k\right\|_2\geq \mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha+1}{c}\log(p)}\right)\mid \mathcal{E}_\alpha^*\right) \leq \frac{C}{p^\alpha}.$$

Thus, using Lemma 6.3 and Lemma 4.2, we obtain

$$\mathbb{P}\left(\max_{k\in T^{*c}}\left\|X_{T^*}^t X_k\right\|_2\geq \frac{C_{col}}{\sqrt{\log(p)}}+\left(1+r+\mathfrak{s}K_{n,s^*}\right)\mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha+1}{c}\log(p)}\right)\mid \mathcal{E}_\alpha^*\right) \leq \frac{C+2}{p^\alpha}.$$

Since, by Assumption (2.7),

$$(1+r+\mathfrak{s}K_{n,s^*})\,\mathfrak{s}\left(\sqrt{n}+\sqrt{\frac{\alpha+1}{c}\log(p)}\right) \quad \leq \quad (1+1.1\cdot r)\,\frac{C_{\mathfrak{s},n,p}}{\sqrt{\log(p)}},$$

we obtain

$$\mathbb{P}\left(\max_{k\in T^{*c}}\left\|X_{T^*}^t X_k\right\|_2 \geq \frac{C_{col}+(1+1.1\cdot r)\;C_{\mathfrak{s},n,p}}{\sqrt{\log(p)}} \mid \mathcal{E}_\alpha^*\right) \quad \leq \quad \frac{C+2}{p^\alpha}.$$

Moreover, using Lemma 4.1, we obtain

$$\mathbb{P}\left(\max_{k\in T^{*c}}\left\|X_{T^*}^t X_k\right\|_2 \geq \frac{C_{col}+(1+1.1\cdot r)\;C_{\mathfrak{s},n,p}}{\sqrt{\log(p)}}\right) \quad \leq \quad \frac{C+3}{p^\alpha}.$$

The last two inequalities

The proof of (3.-15) is standard and, under Assumption 2.7, the proof of (3.-15) can be proved using the ideas of [14, Section 3.3]. We give the proofs for the sake of completeness.

Control of $\left\|X_{T^{*c}}^t X_{T^*}(X_{T^*}^t X_{T^*})^{-1}X_{T^*}^t z\right\|_\infty$

For any $j\in T^{*c}$, we have

$$\begin{aligned}\mathbb{P}\left(X_j^t X_{T^*}(X_{T^*}^t X_{T^*})^{-1}X_{T^*}^t z \geq u\right) \quad &\leq \quad \frac{1}{2}\exp\left(-\frac{u^2}{2\sigma^2\;\left\|X_{T^*}(X_{T^*}^t X_{T^*})^{-1}X_{T^*}^t X_j\right\|_2^2}\right)\\ &\leq \quad \frac{1}{2}\exp\left(-\frac{u^2}{2\sigma^2\;\frac{1+r^*}{(1-r^*)^2}\frac{(C_{col}+(1+1.1\cdot r)\;C_{\mathfrak{s},n,p})^2}{\log(p)}}\right)+\frac{C+219+3}{p^\alpha}\end{aligned}$$

Taking u such that

$$\frac{1}{2}\exp\left(-\frac{u^2}{2\sigma^2\;\frac{1+r^*}{(1-r^*)^2}\frac{(C_{col}+(1+1.1\cdot r)\;C_{\mathfrak{s},n,p})^2}{\log(p)}}\right) \quad = \quad \frac{1}{p^\alpha}$$

i.e.

$$u \quad = \quad \sqrt{(\alpha\log(p)-\log(2))\,2\sigma^2\;\frac{1+r^*}{(1-r^*)^2}\frac{(C_{col}+(1+1.1\cdot r)\;C_{\mathfrak{s},n,p})^2}{\log(p)}}.$$

Using the union bound, we finally obtain

$$\begin{aligned}\mathbb{P}\left(\left\|X_{T^{*c}}^t X_{T^*}(X_{T^*}^t X_{T^*})^{-1}X_{T^*}^t z\right\|_\infty \geq \sqrt{(\alpha\log(p)-\log(2))\,2\sigma^2\;\frac{1+r^*}{(1-r^*)^2}\frac{(C_{col}+(1+1.1\cdot r)\;C_{\mathfrak{s},}}{\log(p)}}\right.\\ \leq \frac{C+223}{p^{\alpha-1}}.\end{aligned}$$

Control of $\left\|X_{T^{*c}}^t X_{T^*}(X_{T^*}^t X_{T^*})^{-1} sign\left(\beta_{T^*}^*\right)\right\|_\infty$

Hoeffding's inequality gives

$$\begin{aligned}\mathbb{P}\left(X_j^t X_{T^*}(X_{T^*}^t X_{T^*})^{-1} sign\left(\beta_{T^*}^*\right) \geq u\right) &\leq \frac{1}{2}\exp\left(-\frac{u^2}{2\ \|(X_{T^*}^t X_{T^*})^{-1} X_{T^*}^t X_j\|_2^2}\right) \\ &\leq \frac{1}{2}\exp\left(-\frac{u^2}{2\ \frac{(C_{col}+(1+1.1\cdot r)\ C_{\mathfrak{s},n,p})^2}{\log(p)\ (1-r^*)^2}}\right) + \frac{C+219+3}{p^\alpha}.\end{aligned}$$

Choosing

$$u = \sqrt{(\alpha\log(p)-\log(2))\,2\ \frac{\left(C_{col}+(1+1.1\cdot r)\ C_{\mathfrak{s},n,p}\right)^2}{\log(p)\ (1-r^*)^2}}.$$

and applying the union bound, we obtain

$$\mathbb{P}\left(X_j^t X_{T^*}(X_{T^*}^t X_{T^*})^{-1} sign\left(\beta_{T^*}^*\right) \geq \sqrt{(\alpha\log(p)-\log(2))\,2\ \frac{\left(C_{col}+(1+1.1\cdot r)\ C_{\mathfrak{s},n,p}\right)^2}{\log(p)\ (1-r^*)^2}}\right) \leq \frac{C+223}{p^{\alpha-1}}.$$

Summing up

Using Assumption 2.6, we obtain that

$$\begin{aligned}&\left\|X_{T^{*c}}^t X_{T^*}(X_{T^*}^t X_{T^*})^{-1} X_{T^*}^t z\right\|_\infty + \lambda\ \left\|X_{T^{*c}}^t X_{T^*}(X_{T^*}^t X_{T^*})^{-1} sign\left(\beta_{T^*}^*\right)\right\|_\infty \\ &\qquad \leq \sigma\sqrt{1+1.1\cdot r\ \ (1.1+0.11\cdot r)} + \frac{1}{2}\ \lambda\end{aligned}$$

as announced.

Bibliography

[1] AlQuraishi, M. and McAdams, H., Direct inference of protein–DNA interactions using compressed sensing methods, PNAS 108, 14819 (2011). (p. 249).

[2] Becker, S., Bobin, J. and Candès, E. J., Nesta: a fast and accurate first-order method for sparse recovery. In press SIAM J. on Imaging Science. (p. 250).

[3] Bickel, P. J., Ritov, Y., Tsybakov, A. B. Simultaneous analysis of lasso and Dantzig selector. Ann. Statist. 37 (2009), no. 4, 1705–1732. (pp. 177 et 250).

[4] Bousquet, O., A Bennett concentration inequality and its application to suprema of empirical processes, Comptes Rendus Mathematique, 334 (2002), no. 6, 495–500. (p. 281).

[5] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007a). Sparsity oracle inequalities for the Lasso. Electron. J. Stat., 1 :169–194. (p. 250).

[6] Bunea, F., Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization , the Electronic Journal of Statistics, (2008) Vol. 2, 1153-1194

[7] Candès, E., Compressive sampling, (2006) 3, International Congress of Mathematics, 1433–1452, EMS. (pp. 167 et 250).

[8] Candès, E. J. The restricted isometry property and its implications for compressed sensing. C. R. Math. Acad. Sci. Paris 346 (2008), no. (p. 250).

[9] Candès, E. J. and Plan, Yaniv. Near-ideal model selection by ℓ_1 minimization. Ann. Statist. 37 (2009), no. 5A, 285–2177. (pp. 177, 178, 180, 181, 182, 183, 184, 185, 187, 188, 189, 192, 193, 205, 206, 207, 217, 218, 219, 220, 228, 229, 230, 233, 243, 245, 250, 254, 255, 256, 286, 300 et 301).

[10] Candès, E. and Tao T., Decoding by linear programming. IEEE Information Theory, 51 (2005) no. 12, 4203-4215. (pp. 250 et 300).

[11] Candès, E. J. and Tao, T., The Dantzig Selector: statistical estimation when p is much larger than n. Ann. Stat. (pp. 177 et 250).

[12] Chrétien, S. and Darses, Sparse recovery with unknown variance: a LASSO-type approach, ArXiv 2012.

[13] Chrétien, S. and Darses, S. Invertibility of random submatrix via tail decoupling and a matrix Chernoff inequality, Stat. and Prob. Lett. 82 (2012), no. 7, 1479-1487. (pp. 229, 245 et 282).

[14] Constructing message passing algorithms for compressed sensing D. L. Donoho, A. Maleki, and A. Montanari, submitted to IEEE Trans. Inf. Theory. (p. 250).

[15] Dossal, C., A necessary and sufficient condition for exact recovery by ℓ_1 minimization. http://hal.archives-ouvertes.fr/docs/00/16/47/38/PDF/DossalMinimisationl1.pdf (p. 250).

[16] Fuchs, J.J., On sparse representations in arbitrary redundant bases. IEEE Trans. Info. Th., 2002. (p. 250).

[17] Kim, S.-J., Koh, K., Boyd, S. and Gorinevsky, D., SIAM Review, problems and techniques section, 51 , (2009), no. 2, 339–360. (p. 249).

[18] Massart, P., Concentration inequalities and model selection. Lectures from the 33rd Summer school on Probability Theory in Saint Flour. Lecture Notes in Mathematics, 1896. Springer Verlag (2007). (p. 281).

[19] Meinshausen, N. and Bühlmann, P., High-dimensional graphs and variable selection with the Lasso, Ann. Statist. 34 (2006), no. 3, 1436-1462. (p. 249).

[20] Neto, D. Sardy, S. and Tseng, P., l1-Penalized Likelihood Smoothing and Segmentation of Volatility Processes allowing for Abrupt Changes, Journal of Computational and Graphical Statistics, 21 (2012), no. 1, 217–233. (p. 249).

[21] Osborne, M.R., Presnell, B. and Turlach, B.A., A new approach to variable selection in least squares problems, IMA J. Numer. Anal. 20 (2000), no. 3, 389–403. (p. 250).

[22] Rudelson, M. and Vershynin, R. Smallest singular value of a random rectangular matrix. Comm. Pure Appl. Math. 62 (2009), no. 12, 1707Ž0131739. (pp. 236 et 280).

[23] Tibshirani, R. Regression shrinkage and selection via the LASSO, J.R.S.S. Ser. B, 58, no. 1 (1996), 267–288. (pp. 177, 228 et 250).

[24] Tropp, J. A. Norms of random submatrices and sparse approximation. C. R. Math. Acad. Sci. Paris 346 (2008), no. 23-24, 1271–1274. (pp. 181, 187, 193, 218, 219, 222, 282 et 300).

[25] Tropp, J. A., User friendly tail bounds for sums of random matrices, http://arxiv.org/abs/1004.4389, (2010). (p. 280).

[26] van de Geer, S., High-dimensional generalized linear models and the Lasso. The Annals of Statistics 36, 614-645.

[27] Vershynin, R., Introduction to the non-asymptotic analysis of random matrices, Chapter 5 of the book Compressed Sensing, Theory and Applications, ed. Y. Eldar and G. Kutyniok. Cambridge University Press, 2012. pp. 210–268. [arXiv:1011.3027, Aug 2010]. (p. 281).

[28] Wainwright, Martin J., Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory 55 (2009), no. 5, 2183–2202.

(pp. 177, 186, 187, 229 et 250).

Chapter L

On the perturbation of the extremal singular values of a matrix after appending a column

with Sébastien Darses.

Abstract

We provide new bounds on the extreme singular values of a matrix obtained after appending a column vector to a given matrix. The proposed bounds improve upon the results obtained in [24]. Moreover, we present two applications of independent interest: a first one regarding the restricted isometry constant and the coherence in Compressed Sensing theory, and a second one concerning the perturbation of the algebraic connectivity of a graph after removing an edge.

1 Introduction

Framework

Let d be an integer. Let $X \in \mathbb{R}^{d\times n}$ be a $d \times n$-matrix and let $x \in \mathbb{R}^d$ be column vector. We denote by a subscript ^{t} the transpose of vectors and matrices. There exist at least two ways to study the matrix (x, X) obtained by appending the column vector x to the matrix X:

(A1) Consider the matrix

$$A \;=\; \begin{bmatrix} x^t \\ X^t \end{bmatrix} \begin{bmatrix} x & X \end{bmatrix} = \begin{bmatrix} x^t x & x^t X \\ X^t x & X^t X \end{bmatrix}; \tag{1.1}$$

(A2) Consider the matrix

$$\widetilde{A} \;=\; \begin{bmatrix} x & X \end{bmatrix} \begin{bmatrix} x^t \\ X^t \end{bmatrix} = XX^t + xx^t.$$

On one hand, one may study in (A1) the eigenvalues of the $(n+1)\times(n+1)$ hermitian matrix A, i.e. the matrix X^tX augmented with an arrow matrix.

On the other hand, one will deal in (A2) with the eigenvalues of the $d \times d$ hermitian matrix $\widetilde{A}$, which may be seen as a rank-one perturbation of XX^t. The matrices A and $\widetilde{A}$ have the same non-zeros eigenvalues, and in particular $\lambda_{\max}(A) = \lambda_{\max}(\widetilde{A})$. Moreover, the singular values of the matrix (x, X) are the square-root of the eigenvalues of the matrix A.

Equivalently, the problem of a rank-one perturbation can be rephrased as the one of controlling the perturbation of the singular values of a matrix after appending a column.

In the current paper, we study a slightly more general framework than (A1), that is the case of a matrix

$$A = \begin{bmatrix} c & a^t \\ a & M \end{bmatrix}, \tag{1.1}$$

where $a \in \mathbb{R}^d$, $c \in \mathbb{R}$ and $M \in \mathbb{R}^{d\times d}$ is a symmetric matrix.

Our goal is to present new bounds on the extreme eigenvalues of A as a function of the eigenvalues of M and the norm of a, and we will focus on various applications. Indeed, this problem occurs in a variety of contexts such as the perturbation analysis of covariance matrices in statistics [28], the study of the Restricted Isometry Constant in Compressed Sensing [9], spectral graph theory and edge deletion [8], control theory of complex networks [31], hitting time analysis for classical or quantum random walks [38], robust face recognition [32], wireless comunications [34], communication theory and signal processing [38], numerical methods for partial differential equations [4], numerical analysis of bifurcations [17], among many applications.

Notice further that in (1.1) if M and A are positive definite, there exist $X \in \mathbb{R}^{d\times n}$ and $x \in \mathbb{R}^d$ such that $M = X^tX$ and A can be written as in (1.1) due to the Chowlesky decomposition.

Additional notations

The Kronecker symbol is denoted by $\delta_{i,j}$, i.e. $\delta_{i,j} = 1$ if $i = j$ and is equal to zero otherwise. For any symmetric matrix $B \in \mathbb{R}^{d\times d}$ we will denote its eigenvalues by $\lambda_1(B) \geq \cdots \geq \lambda_d(B)$. The largest eigenvalue will sometimes also be denoted by $\lambda_{\max}(B)$ and the smallest by $\lambda_{\min}(B)$. The smallest nonzero eigenvalue of B will be denoted by $\lambda_{\min>0}(B)$.

Plan of the paper

Section 2 is devoted to an overview of known results. Section 3 presents new upper and lower bounds for the extreme eigenvalues. Section translates some previous results in terms of operator norm together with a slight variation. Finally, Section 5 is concerned with the applications in Compressed sensing and graphs theory.

2 Previous results on eigenvalue perturbation

We now review some previous, old and recent results from matrix perturbation theory and apply them to our problem of appending a column.

Obtaining precise estimates on the eigenvalues of a sum of two matrices (say $X + P$, considering P as a perturbation) is a very difficult task in general. Weyl's and Horn's inequalities for instance can be employed and these bounds can be improved when knowing that the perturbation P is small with respect to X (see e.g. [20, Chap. 6]). The whole point of the works [2] and [3], to name a few, is to understand how randomness can simplify this analysis.

Weyl's inequalities

The reference [35] gives an overview of many inequalities on the eigenvalues of sums of symmetric (and Hermitian) matrices. The Weyl inequalities are given as follows:

Theorem 2.1 (Weyl) *Let B and B' be symmetric real matrices in $\mathbb{R}^{d\times d}$ and let $\lambda_j(B)$, $j=1,\ldots,d$, (resp. $\lambda_j(B')$), denote the eigenvalues of B (resp. B'). Then, we have*

$$\lambda_{i+j-1}(B+B') \quad \leq \quad \lambda_i(B)+\lambda_j(B'),$$

whenever $i,j\geq 1$ and $i+j-1\leq n$.

The arrowhead perturbation

Consider the case where we would like to control the largest eigenvalue of A with the eigenvalues of $M=X^tX$. We have the following result.

Proposition 2.2 *We have*

$$\lambda_1(A) \quad \leq \quad \max\{c,\lambda_1(M)\}+\|a\|_2.$$

extbfProof. The Weyl inequalities for $i,j=1$ gives that

$$\lambda_1(A) \quad \leq \quad \lambda_1\left(\begin{bmatrix} c & 0 \\ 0 & M \end{bmatrix}\right)+\lambda_1(E) \tag{2.0}$$

with

$$E=\begin{bmatrix} 0 & a^t \\ a & 0 \end{bmatrix}.$$

Moreover, using the variational representation of the maximum eigenvalue and the method of Lagrange multipliers, we have $\lambda_1(E)=\|a\|_2$. Combining this with (2.0), we obtain the desired result. □ The main fact to retain from this inequality is that if x is orthogonal to all columns of X, then $a=0$ and the perturbation has no effet on the largest eigenvalue as long as $c\leq\lambda_1(M)$. This elementary observation can be extrapolated to much more difficult situations, e.g. in the spiked covariance model where a phase transition has been proved between concerning the ability to detect a spike or not, depending on the energy level of the spike [28, Theorem 2.3].

The rank-one perturbation

If we only want to study the perturbation of the largest eigenvalue, then we can consider the rank-one perturbation described by (A2). In this case, Weyl's bound gives the following result.

Proposition 2.3 *We have*

$$\lambda_1(A) \quad \leq \quad \lambda_1(M)+\|x\|_2^2.$$

extbfProof. Using that $\lambda_1(A)=\lambda_1(\widetilde{A})$ and $\lambda_1(M)=\lambda_1(\widetilde{M})$, we obtain from Theorem 2.1 :

$$\lambda_1(A) \quad \leq \quad \lambda_1(M)+\lambda_1(xx^t).$$

Since $\lambda_1(xx^t)=\|x\|_2^2$, the conclusion follows. □
The main drawback of this inequality is that it does not take into account the geometry of the problem and in particular the angle between X and the new vector x that we want to append to X. This does not disqualify the rank-one perturbation approach to controlling the maximum eigenvalue as will be shown in Subsection 2.

An inequality of Li and Li

They prove a general inequality concerning the perturbation of eigenvalues under off-block diagonal perturbations. We specify their result, [24, Theorem 2], in our context:

$$|\lambda_1(A) - \max(c, \lambda_1(M))| \quad \leq \quad \frac{2\|a\|^2}{\eta + \sqrt{\eta^2 + 4\|a\|^2}}, \tag{2.-2}$$

with $\eta = \min\{|c - \lambda_i(M)|, 1 \leq i \leq d\}$. In their paper, $\widetilde{\lambda}_1$ is actually $\max(c, \lambda_1(M))$ here.

We refer to [24] and references therein for the history of such inequalities.

An inequality of Ipsen and Nadler

In [21], the authors propose a bound for the eigenvalues of $\widetilde{A}$ in the problem of rank one perturbation (A2). The following theorem is a corollary of their main result where we restrict our attention to the largest eigenvalue.

Theorem 2.4 *Let $\widetilde{M} \in \mathbb{C}^{d\times d}$ denote an Hermitian matrix and let $x \in \mathbb{C}^d$. Let V_1 (resp. V_2) denote the eigenvector associated to the eigenvalue $\lambda_1(\widetilde{M})$ (resp. $\lambda_2(\widetilde{M})$). Let $\widetilde{A} = \widetilde{M} + xx^t$. Then*

$$\lambda_1(\widetilde{M}) + \delta_{\min} \quad \leq \lambda_1(\widetilde{A}) \leq \quad \lambda_1(\widetilde{M}) + \delta_{\max},$$

with

$$\begin{aligned}
\delta_{\min} &= \frac{1}{2}\left(\|P_{\langle (V_1,V_2)\rangle}(x)\|_2^2 - gap_2 + \sqrt{(gap_2 + \|P_{\langle (V_1,V_2)\rangle}(x)\|_2^2)^2 - 4\ gap_2\|P_{\langle (V_2)\rangle}(x)\|_2^2}\right)\\
\delta_{\max} &= \frac{1}{2}\left(\|x\|_2^2 - gap_2 + \sqrt{(gap_2 + \|x\|_2^2)^2 - 4\ gap_2\|P_{\langle (V_2,\ldots,V_d)\rangle}(x)\|_2^2}\right),
\end{aligned}$$

where $(V_i, \ldots, V_j)$, $1 \leq i \leq j \leq d$, denotes the vector space generated by $V_i, \ldots, V_j$ and $P_{\langle (V_i,\ldots,V_j)\rangle}$ denotes the orthogonal projection onto this space, and

$$gap_2 \quad = \quad \lambda_1(\widetilde{M}) - \lambda_2(\widetilde{M}).$$

This inequality has been used in various applications such as control of complex systems [31], quantum information theory [15], communication theory and signal processing [38], numerical methods for partial differential equations [4]. One drawback of using this result in our context is that we have to know the spacing gap_2 for the second eigenvalue. Moreover, the upper bound depends on $\|x\|_2^2$ and does not take into account the scalar products of x with the columns of X, which may lead to serious overestimation of the perturbation, especially in the case of random matrices.

3 Main results on the perturbation of the extreme singular values

In this section, we present and prove our main results. We improve the bound obtained from Weyl's inequality over a non-trivial and useful range of perturbations. Moreover, our bound does not depend on the spacing gap_2 unlike in [21].

The maximum eigenvalue

The following theorem provides sharp upper bounds for $\lambda_{\max}(A)$, and lower bounds on $\lambda_{\min}(A)$, depending on various informations on the sub-matrix M of A. As discussed above, this problem has close relationships with our problem of appending a column to a given rectangular matrix, because $\lambda_1(\widetilde{A}) = \lambda_1(A)$.

Theorem 3.1 *Let d be a positive integer and let $M \in \mathbb{C}^{d\times d}$ be an Hermitian matrix, whose eigenvalues are $\lambda_1 \geq \cdots \geq \lambda_d$ with corresponding eigenvectors $(V_1, \cdots, V_d)$. Set $c \in \mathbb{R}$, $a \in \mathbb{C}^d$. Let A be given by (1.1). Therefore:*

$$\frac{2\langle a, V_1\rangle^2}{\eta_1 + \sqrt{\eta_1^2 + 4\langle a, V_1\rangle^2}} \leq \lambda_1(A) - \max(c, \lambda_1) \leq \frac{2\|a\|^2}{\eta_1 + \sqrt{\eta_1^2 + 4\|a\|^2}}, \tag{3.-6}$$

with

$$\eta_1 \quad = \quad |c - \lambda_1|.$$

Remark 3.2

- *Inequality (3.1) is sharp: the upper bound is reached when choosing $M = I$, $c = 1$ and any a, so that $\lambda_{\max}(A) = 1 + \|a\|$;*
- *The upper bound in (3.1) is better than (2.-2) since $\eta_1 \geq \eta$. A typical example where the improvement holds is basically when c is one of the eigenvalues of M (i.e. $\eta = 0$). For instance, take $c = 1$, $a^t = (\alpha, 0)$ and $M = diag(2, 1)$. In particular, $\eta_1 = 1$. An easy computation yields $\lambda_1(A) = 3/2 + \sqrt{1/4 + \alpha^2}$ and then*

$$\lambda_1(A) - \lambda_1(M) = \sqrt{1/4 + \alpha^2} - 1/2 = \frac{2\alpha^2}{1 + \sqrt{1 + 4\alpha^2}},$$

  which is the upper bound in (3.1), while the bound (2.-2) is simply the triangle inequality $|\lambda_1(A) - \lambda_1(M)| \leq |\alpha|$.

- *The lower bound in (3.1) is also better than (2.-2) since we have:*

$$\lambda_1(A) \geq \max(c, \lambda_1) + \frac{2\langle a, V_1\rangle^2}{\eta_1 + \sqrt{\eta_1^2 + 4\langle a, V_1\rangle^2}} \geq \max(c, \lambda_1) - \frac{2\|a\|^2}{\eta + \sqrt{\eta^2 + 4\|a\|^2}}.$$

  Our lower bound is in particular consistent with Cauchy interlacing theorem, which states that $\lambda_1(A) \geq \lambda_1$.

- *A great feature of Theorem 2 of Li and Li in [24] is that it holds for all eigenvalues and for block perturbations. We will treat the whole spectrum in a subsequent work. Moreover, generalizing our bounds for block perturbations may be an interesting perspective.*

Proof.

Let $M = VDV^*$ denote the eigenvalue decomposition of M, i.e. $V = (V_1, \cdots, V_d)$ where the V_i's are the orthonormal eigenvectors of M and D is a diagonal matrix whose diagonal entries are the real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. We can write

$$A \quad = \quad \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} c & a^*V \\ V^*a & D \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & V^* \end{pmatrix},$$

and we set

$$B = \begin{pmatrix} c & b^* \\ b & D \end{pmatrix}, \quad b = V^* a,$$

where we use the notation $b_j := \langle a, V_j \rangle$. Therefore, A and B have the same spectra and in particular,

$$\lambda_{\max}(A) = \lambda_{\max}(B). \tag{3.-10}$$

As in [16], we compute the characteristic polynomial of the arrow matrix B:

$$P_B(\lambda) = (c-\lambda)\prod_{i=1}^{d}(\lambda_i - \lambda) - \sum_{i=1}^{d}\prod_{j\neq i}(\lambda_j - \lambda)b_j^2.$$

Let us define the function f on $\mathbb{R} \setminus \{\lambda_i, 1 \leq i \leq d\}$ as

$$f(\lambda) := P_B(\lambda)\prod_{i=1}(\lambda_i - \lambda)^{-1} = c - \lambda + \sum_{j=1}^{d}\frac{b_j^2}{\lambda - \lambda_j},$$

which is decreasing on $(\lambda_1, +\infty)$ (even if $b = 0$).

We now assume that $b_1 = \langle a, V_1 \rangle \neq 0$. Thus $\lim_{\lambda\to\lambda_1} f(\lambda) = +\infty$. From $\lim_{\lambda\to+\infty} f(\lambda) = -\infty$, we then deduce that the continuous function f has a unique root on $(\lambda_1, +\infty)$, that is

$$\lambda_{\max}(B) > \lambda_1.$$

For all $\lambda > \lambda_1$, we have

$$f(\lambda) \leq c - \lambda + \frac{\|b\|_2^2}{\lambda - \lambda_1} := g(\lambda). \tag{3.-12}$$

For the same reasons as f, the function g has a unique root λ^* on $(\lambda_1, +\infty)$. Since f is decreasing on $(\lambda_1, +\infty)$ and $f(\lambda_{\max}(B)) = 0 = g(\lambda^*) \geq f(\lambda^*)$, we deduce:

$$\lambda_{\max}(B) \leq \lambda^*.$$

We have

$$(\lambda^* - c)(\lambda^* - \lambda_1) = \|b\|_2^2,$$

and thus λ^* is a root of the polynomial

$$\begin{aligned} Q(x) &= (x-c)(x-\lambda_1) - \|b\|^2 \\ &= x^2 - (c+\lambda_1)x + c\lambda_1 - \|b\|_2^2. \end{aligned}$$

The discrimant of Q reads:

$$\begin{aligned} \Delta &= (c+\lambda_1)^2 - 4(c\lambda_1 - \|b\|_2^2) \\ &= (c-\lambda_1)^2 + 4\|b\|_2^2 > 0. \end{aligned}$$

Since $Q(\lambda_1) < 0$ and the dominant coefficient of Q is positive, we deduce that λ^* is actually the greatest root of Q. Hence, noting that $\|b\|_2 = \|a\|_2$,

$$\lambda^* = \frac{c+\lambda_1}{2} + \frac{1}{2}\sqrt{(c-\lambda_1)^2 + 4\|a\|^2}. \tag{3.-17}$$

Assume that $\langle a, V_1\rangle \neq 0$. In order to find a lower bound for $\lambda_{\max}(B)$, we perform the same reasoning by writing

$$f(\lambda) \quad \geq \quad c - \lambda + \frac{\langle a, V_1\rangle^2}{\lambda - \lambda_1},$$

and considering the polynomial $(x-c)(x-\lambda_1) - \langle a, V_1\rangle^2$.

Finally, we have:

$$\frac{c+\lambda_1}{2} + \frac{1}{2}\sqrt{(c-\lambda_1)^2 + 4\langle a, V_1\rangle^2} \leq \lambda_1(A) \leq \frac{c+\lambda_1}{2} + \frac{1}{2}\sqrt{(c-\lambda_1)^2 + 4\|a\|^2}. \qquad (3.\text{-}18)$$

Set $\eta_1 = |c - \lambda_1|$. Since

$$2\max(\lambda_1, c) \quad = \quad c + \lambda_1 + \eta_1,$$

we deduce

$$\frac{1}{2}\left(\sqrt{\eta_1^2 + 4\langle a, V_1\rangle^2} - \eta_1\right) \leq \lambda_1(A) - \max(\lambda_1, c) \leq \frac{1}{2}\left(\sqrt{\eta_1^2 + 4\|a\|^2} - \eta_1\right).$$

Multiplying by the "conjugate quantity" yields the lower and the upper bounds in (3.1).

The case $\langle a, V_1\rangle = 0$ can be treated by standard continuity arguments: consider a continuous $\varepsilon \mapsto a(\varepsilon)$ such that for all $\varepsilon > 0$, $\langle a(\varepsilon), V_1\rangle \neq 0$ and $a(0) = a$. Ones then writes (3.1) for $\varepsilon > 0$ and passes to the limit as $\varepsilon \to 0$.

□

Corollary 3.3 *In particular, the following simple perturbation bounds hold:*

$$\lambda_1(A) \quad \leq \quad \max(c, \lambda_1) + \|a\|_2 \qquad (3.\text{-}19)$$

$$\lambda_1(A) \quad \leq \quad \max(c, \lambda_1) + \frac{\|a\|_2^2}{|\lambda_1 - c|}, \qquad (3.\text{-}18)$$

extbfProof. Inequality (3.-19) (resp. (3.-18)) follows from (3.1) by using $\eta_1 \geq 0$ (resp. $\|a\| \geq 0$).

Perturbation of the smallest nonzero eigenvalue

Theorem 3.4 *Let d be a positive integer and let $M \in \mathbb{C}^{d\times d}$ be an Hermitian matrix, whose eigenvalues are $\lambda_1 \geq \cdots \geq \lambda_d$ with corresponding eigenvectors $(V_1, \cdots, V_d)$. Set $c \in \mathbb{R}$, $a \in \mathbb{C}^d$. Let A be given by (1.1). Assume that M has rank $r \leq d$. Therefore:*

$$\lambda_{r+1}(A) \geq \min(c, \lambda_r) - \frac{2\|a\|^2}{\eta_r + \sqrt{\eta_r^2 + 4\|a\|^2}}, \qquad (3.\text{-}18)$$

with

$$\eta_r \quad = \quad |c - \lambda_r|.$$

extbfProof.

In the case where $r = d$, Inequality (3.4) immediately follows from applying (3.-19) to the matrix $-A$.

Assume now that $r < d$. The eigenvalues of M are $\lambda_1 \geq \cdots \geq \lambda_r \geq \lambda_{r+1} = \cdots = \lambda_d = 0$. We use the same reduction for the matrix A as in the proof of Theorem 3.1. If $a \neq 0$, we have that $b \neq 0$ and so there exists j_0 such that $b_{j_0} \neq 0$. Again, we consider the function f, defined on $\mathbb{R} \setminus \{\lambda_i, 1 \leq i \leq d\}$ by

$$f(\lambda) \quad := \quad c - \lambda + \sum_{j=1}^{d} \frac{b_j^2}{\lambda - \lambda_j},$$

which is decreasing on $(\lambda_{r+1}, \lambda_r)$. Since $\lim_{\lambda \downarrow \lambda_{r+1}} f(\lambda) = +\infty$ and $\lim_{\lambda \uparrow \lambda_r} f(\lambda) = -\infty$, we deduce that the continuous function f has a unique root on $(\lambda_{r+1}, \lambda_r)$, which is $\lambda_{r+1}(A)$. We have

$$f(\lambda) \quad := \quad c - \lambda + \sum_{j=1}^{r} \frac{b_j^2}{\lambda - \lambda_j} + \sum_{j=r+1}^{d} \frac{b_j^2}{\lambda},$$

For all λ s.t. $0 = \lambda_{r+1} < \lambda < \lambda_r$, we have

$$f(\lambda) \quad \geq \quad c - \lambda + \frac{\sum_{j=1}^{r} b_j^2}{\lambda - \lambda_r} + \frac{\sum_{j=r+1}^{d} b_j^2}{\lambda} \tag{3.-20}$$

and thus

$$f(\lambda) \quad \geq \quad c - \lambda + \frac{\sum_{j=1}^{d} b_j^2}{\lambda - \lambda_r} := g(\lambda). \tag{3.-19}$$

The function g has a unique root λ^* in $(-\infty, \lambda_r)$.

Since f is decreasing on $(0 = \lambda_{r+1}, \lambda_r)$ and $f(\lambda_{r+1}(B)) = 0 = g(\lambda^*) \leq f(\lambda^*)$, we deduce that:

$$\lambda_{r+1}(A) \quad \geq \quad \lambda^*. \tag{3.-18}$$

Let us now bound λ^* from below. We have

$$(c - \lambda^*)(\lambda_r - \lambda^*) = \sum_{j=1}^{d} b_j^2 = \|b\|_2^2 = \|a\|_2^2. \tag{3.-17}$$

Therefore

$$\lambda^{*^2} - (c + \lambda_r)\lambda^* + c\ \lambda_r - \|a\|_2^2 \quad = \quad 0. \tag{3.-16}$$

and thus,

$$\lambda^* \quad = \quad \frac{1}{2}\left(c + \lambda_r - \sqrt{(c + \lambda_r)^2 - 4c\ \lambda_r + 4\|a\|_2^2}\right), \tag{3.-15}$$

since one easily checks that the other root is greater than λ_r. Expanding the term $(c + \lambda_r)^2$ inside the square root and simplifying the resulting expression, we get

$$\lambda^* \quad = \quad \frac{1}{2}\left(c + \lambda_r - \sqrt{(c - \lambda_r)^2 + 4\|a\|_2^2}\right) \tag{3.-14}$$

and

$$\lambda^* - \min(c, \lambda_r) \quad \geq \quad \frac{1}{2}\left(\eta_r - \sqrt{\eta_r^2 + 4\|a\|_2^2}\right), \tag{3.-13}$$

with

$$\eta_r \quad = \quad |c - \lambda_r|.$$

The desired result then follows by multiplying by the "conjugate quantity".
□

Corollary 3.5 *In particular, the following simple perturbation bounds hold:*

$$\lambda_{r+1}(A) \quad \geq \quad \min(c, \lambda_r) - \|a\|_2 \tag{3.-13}$$

$$\lambda_{r+1}(A) \quad \geq \quad \min(c, \lambda_r) - \frac{\|a\|_2^2}{|c - \lambda_r|}. \tag{3.-12}$$

extbfProof. Inequality (3.-13) (resp. (3.-12)) follows from (3.4) by using $\eta_r \geq 0$ (resp. $\|a\| \geq 0$).

4 Bounds on the perturbation of the operator norm

We provide here three bounds on the operator norm: the first and second inequalities are easy consequences of Theorem 3.1, the third one is based on a new trick.

Corollary 4.1 *Let d be an integer, $a \in \mathbb{C}^d$, $c \in \mathbb{R}$ and let $M \in \mathbb{C}^{d \times d}$ be an Hermitian matrix. Let A be given by (1.1). Then the following inequalities hold:*

$$\|A\| \quad \leq \quad \max(c, \|M\|) + \|a\|_2 \tag{4.-11}$$

$$\|A\| \quad \leq \quad \|M\| + \frac{\|a\|_2^2}{\|M\| - c}, \quad if\ c \leq \lambda_{\max}(M) \tag{4.-10}$$

$$\|A\| \quad \leq \quad \|M\| + \frac{|c|}{2} + \frac{\|a\|_2^2 + c^2/8}{\|M\|}. \tag{4.-9}$$

Remark 4.2 *Notice that (4.-10) is better than (4.-11) if*

$$\|a\| \quad \leq \quad \|M\| - c,$$

and that (4.-9) is better than (4.-11) if

$$\frac{c}{2} + \frac{\|a\|_2^2 + c^2/8}{\|M\|} \quad \leq \quad \|a\|.$$

extbfProof. We obtain (4.-11) by applying (3.-19) with $-A$ and by noticing that $\lambda_{\max}(A) \leq \|A\|$.

Now assume that $c \leq \lambda_{\max}(M)$. We bound Δ as:

$$\sqrt{\Delta} \leq \sqrt{(\|M\| - c)^2 + 4\|a\|_2^2},$$

and then

$$2\lambda^* \quad \leq \quad 2\|M\| + \frac{2\|a\|_2^2}{\|M\| - c},$$

which yields (4.-10).

To prove (4.-9), we now consider, instead of B,

$$B' = \begin{pmatrix} c & b^t & b^t \\ b & D & 0 \\ b & 0 & -D \end{pmatrix}. \tag{4.-12}$$

Since the operator norm increases by adding elements to a matrix , we obtain

$$\|A\| \leq \|B'\| \tag{4.-11}$$

The functions f, g in (3.-20) are now replaced resp. by,

$$\begin{aligned} \widetilde{f}(\lambda) &= c - \lambda + \sum_{j=1}^{d} b_j^2 \left(\frac{1}{\lambda - \lambda_j} + \frac{1}{\lambda + \lambda_j} \right) = c - \lambda + \sum_{j=1}^{d} b_j^2 \, \frac{2\lambda}{\lambda^2 - \lambda_j^2} \\ \widetilde{g}(\lambda) &= c - \lambda + \|b\|_2^2 \, \frac{2\lambda}{\lambda^2 - \|M\|^2}, \quad \lambda > \|M\|. \end{aligned}$$

If $c \leq 0$ then

$$\widetilde{f}(\lambda) \leq \widetilde{g}(\lambda) \leq \lambda + \|b\|_2^2 \, \frac{2\lambda}{\lambda^2 - \|M\|^2} := h(\lambda).$$

Let x^* be a root of h. As previously, $\widetilde{f}(\lambda_{\max}(\widetilde{B})) = 0 = h(x^*) \geq \widetilde{f}(x^*)$, and then

$$\lambda_{\max}(\widetilde{B}) \leq x^*.$$

But x^* is less than the greatest root of the polynomial $x \mapsto x^2 - \|M\|^2 + 2\|b^2\|$, that is:

$$x^* \leq \sqrt{\|M\|^2 + 2\|b\|_2^2}.$$

If $c > 0$, we notice that

$$\begin{aligned} (\lambda^2 - \|M\|^2)(c - \lambda) + 2\lambda\|b\|_2^2 &= -\lambda^3 + c\lambda^2 + (2\|b\|_2^2 + \|M\|^2)\lambda - c\|M\|^2 \\ &\leq -\lambda^3 + c\lambda^2 + (2\|b\|_2^2 + \|M\|^2)\lambda, \end{aligned}$$

and we set

$$R(x) = x^2 - cx - (2\|b\|_2^2 + \|M\|^2).$$

The greatest root x^* of R reads:

$$\begin{aligned} x^* &\leq \frac{c}{2} + \sqrt{\frac{c^2}{4} + \|M\|^2 + 2\|b\|_2^2} \\ &\leq \frac{c}{2} + \|M\| \sqrt{1 + \frac{2\|b\|_2^2 + c^2/4}{\|M\|^2}} \\ &\leq \frac{c}{2} + \|M\| + \frac{\|b\|_2^2 + c^2/8}{\|M\|}. \end{aligned}$$

Repeating the analysis with $-A$ yields (4.-9) as desired. □

5 Applications

As already mentionned in the introduction, perturbations bounds on the extreme eigenvalues have many applications in science and engineering and some references were proposed. In this section, we focus two more applications where quadratic inequalities as the upper bound (3.1) can yield some improvements in the order of magnitude for the perturbed system.

Restricted isometry constant and coherence in Compressed Sensing

General framework

The purpose of Compressed Sensing (CS) is to study the various possible strategies for constructing efficient sensors allowing the recovery of very sparse signals in a high dimensional space (See e.g. the pioneering work of Candès, Romberg and Tao [13]). The possiblity of building such types of sensors was first discovered through simulations in the study of Magnetic Resonnance Imaging, where sparsity in a certain dictionary was used in order to reconstruct the signal from much fewer measurements than was previously imagined. Since then, Compressed Sensing has found many applications as can be seen from the blog "Nuit Blanche" maintained by Igor Caron.

The problem can be expressed mathematically as the one of solving the linear system

$$y \quad = \quad X\beta + \sigma\varepsilon$$

in the variable β, where $X \in \mathbb{R}^{n\times p}$, $\sigma \in \mathbb{R}_+$ and ε is a random noise. A major breakthrough occured in late 2005-early 2006 when [13], [12], [11] and [10] appeared. One of the main discoveries contained in these works is that the vector β can be recovered exactly even when p is much larger than n and n is as small as a constant times $s\log(p/s)$. The assumptions initially required that $\sigma = 0$ and β is s-sparse and the results were obtained for most X drawn with i.i.d. components with standard gaussian or ± 1-Bernoulli distribution. It was then obtained in [11] and [14] that the support of β can be exactly recovered in the noisy case $\sigma > 0$ when n is roughly of the same order. A basic property, which emerged from the analysis as a tool for proving the reconstructibility from few measurements, is the Restricted Isometry Property, which requires that all the submatrices X_T have their singular values in the interval $[1-\rho, 1+\rho]$ for some constant $\rho \in (0, 1/2)$. Several authors [36], [37] and [14] subsequently noticed that, assuming the columns of X to be ℓ_2-normalized, most submatrices X_T obtained by selecting the columns indexed by T with $|T|$ such that

$$|T| \quad \leq \quad \frac{p}{\log p}\,\frac{C}{\|X\|^2} \tag{5.-22}$$

for some constant C, have their singular values in the interval $[1-\rho, 1+\rho]$ for some constant $\rho \in (0, 1/2)$. Recall that the coherence $\mu(X)$ is defined by

$$\mu(X) \quad = \quad \max_{j\neq j'} |X_j^t X_{j'}|.$$

This latter property can be interpreted in a probabilistic setting: let T be a random subset of $\{1, \ldots, n\}$ drawn with uniform distribution over all subsets with cardinal bounded from above as in (5.-22). Then, with high probability, $\|X_T^t X_T - I\| \leq \rho$.

Perturbation of the singular values

When an additional column is appended to the matrix X, one may wonder what is the impact of this operation on the localisation of the extreme singular values of all submatrices

with s columns which can be extracted from the resulting matrix. Notice that appending just one column to X results in creating $p!/(s-1)!(p-s+1)!$ additional submatrices. Therefore, having a flexible bound on the perturbation of the extreme eigenvalues may be a valuable tool in practice. Another situation where perturbation has to be precisely controlled is when one wants to study the random variable $\|X_T^t X_T - I\|$ using the tools of modern concentration of measure theory [6]. Indeed, after a 'Poissonization' trick has been employed as in Claim (3.29) p.2173 in [14], one may study the problem on a product space for which the celebrated theorem of Talagrand or recent variants by Boucheron, Lugosi and Massart can be used. However, for such concentration theorems to be relevant, one also needs precise perturbation bounds on the extreme singular values.

Let us consider the case where one uses a fixed design matrix X and T is obtained by selecting s columns uniformly at random. Then, Lemma 3.6 in [14] implies that

$$\mathbb{P}\left(\|X_T^t X_j\|_2^2 \geq s/p\|X\|^2 + t\right) \quad \leq \quad 2\exp\left(\frac{t^2}{2\mu^2(X)(s\|X\|^2/p + t/3)}\right)$$

and thus, using (5.-22), one easily obtains that

$$\|X_T^t X_j\|_2^2 \leq \frac{1}{4\log(p)} \tag{5.-23}$$

with probability at least $1 - 2e^{-\frac{3}{64\mu^2(X)\log(p)}}$ if $C \leq 1/8$. Assuming that the coherence is of the order of $1/\log(p)$, one obtains that (5.-23) holds with high probability. Thus, using inequality (3.-19), one obtains a perturbation of the order of $\log(p)^{-1/2}$ of the maximum eigenvalue of $X_T^t X_T$. On the other hand, if one is interested in the perturbation with norm already larger than $\sqrt{1+\rho}$, (3.-18) gives a perturbation of the norm of the order $\rho^{-1}\log(p)^{-1}$ which is significantly smaller and, as one might check in the assumptions of Theorem 5 in [5], is the right order of magnitude for obtaining the desired concentration of measure for this problem.

Perturbation of the algebraic connectivity of a graph by removing an edge

with Sébastien Darses.

Another application of spectral perturbation is in hypergraph theory.

The Laplacian of a graph

The $G = (V, E)$ denote an oriented graph with vertex set V and edge set E. In such a graph, each edge e has a positive end and a negative end. We say that two vertices are adjacent if they are ends of the same edge. The indicence matrix $\mathcal{I}_G$ associated to G is the matrix whose rows are indexed by the vertices and the columns are indexed by the oriented edges. The (i,j)-entry of $\mathcal{I}_G$ is

$$\mathcal{I}_G(i,j) \quad = \quad \begin{cases} +1 \text{ if vertex } i \text{ is the positive end of edge } j \\ -1 \text{ if vertex } i \text{ is the negative end of edge } j \\ 0 \text{ otherwise.} \end{cases}$$

The adjacency matrix $\mathcal{A}_G$ is the matrix whose rows and columns are indexed by the vertices. The (i, i')-entry of $\mathcal{A}_G$ is

$$\mathcal{A}_G(i,i') \quad = \quad \begin{cases} +1 \text{ if vertex } i \text{ and vertex } i' \text{ are adjacent} \\ 0 \text{ otherwise.} \end{cases}$$

The degree vector of G is the vector d_G where $d_G(i)$ is the number of edges of G to which vertex i is an end. The Laplacian matrix of G is the matrix $\mathcal{L}_G$ defined by

$$\mathcal{L}_G \quad = \quad D(d_G) - \mathcal{A}_G,$$

and the following well known identity holds

$$\mathcal{L}_G \quad = \quad \mathcal{I}_G \mathcal{I}_G^t. \tag{5.-25}$$

If G is not oriented, the degree vector and the adjacency matrix are defined in exactly the same way and any arbitrary orientation of the edges of G will of course provide the same result. Notice that $\mathcal{L}_G$ is positive semi-definite and that 0 is always an eigenvalue of $\mathcal{L}_G$. If the second smallest eigenvalue is nonzero, then the graph G is connected. This second smallest eigenvalue is very important for the study of various graphs and is called the algebraic connectivity of G or Fiedler's value of G. We will denote the algebraic connectivity by $a(G)$. The eigenvalues of the Laplacian of a graph have been the subject of intense research for many years and is connected to various fields of pure and applied mathematics like expander families [19], geometry of Banach spaces [1], Markov chains [7], clustering [25], to name just a few.

Edge deletion and the algebraic connectivity

We now turn to the problem of controling the impact of deleting an edge on the algebraic connectivity of $\mathcal{L}$. The complement of a graph is the graph obtained by putting an edge between every non-adjacent couple of vertices and by deleting all edges already present in the graph before this operation. It is well known [26] that

$$a(G) \quad \geq \quad n - \lambda_1(G^c). \tag{5.-24}$$

Thus, controlling the effect of adding an edge to the complement of a graph allows to control the effect of deleting an edge of the graph on the algebraic connectivity.

For $e = (u, v)$, with $u, v \in V(G)$, let $G^c + e$ denote the graph obtained from G^c by appending the edge e. Let i_e denote the column vector obtained by setting the component indexed by u to -1 and the component indexed by v to $+1$, and by setting all other components to zero. Since the Laplacian matrix $\mathcal{L}_{G^c}$ admits a factorization analogous to (5.-25), we obtain that $\mathcal{L}_{G^c}$ can be written in the form (1.1) with $c = 2$ and $a = \mathcal{I}_{G^c}^t i_e$.

In many fields, it is very important to study the robustness of the graph topology to structural perturbations. For instance, the study of food webs has been of growing interest in the recent years [33]. As is well known, predation habits evolve with time as a consequence of landscape changes and competition. The world wide web is also an interesting application of graph theory and the formation and perturbation of communities is a topic of growing interest [29]. Communication systems are also often viewed as an interesting application of graph theory. In these examples, as in many other from ecology, social sciences, wireless communications, genetics, etc, one is often interested in predicting the impact on topology of removing or adding an edge, a vertex or of various other modifications of the structure, as measured by a relevant index such as the algebraic connectivity.

Controllability of complex networks

In [31], the following model was proposed. One considers a set of N n-dimensional oscillators governed by a system of nonlinear differential equations. Moreover, we assume that each oscillator is coupled with a restricted set of other oscillators. This coupling relationship can be efficiently described using a graph where the vertices are indexed by the oscillators and there is an edge between two oscillators if they are coupled. The overall dynamical system is given by the following set of differential equations

$$x_i'(t) \quad = \quad f(x_i(t)) - \sigma B \sum_{j=1}^{N} l_{ij} x_j(t) + u_i(t), \ t \geq t_0, \tag{5.-23}$$

$i = 1, \ldots, N$, where $x_i(t) \in \mathbb{R}^n$ is the state of the i^{th} oscillator, σ is a positive real number, $B \in \mathbb{R}^{n \times n}$, $f : \mathbb{R} \mapsto \mathbb{R}$ describes the dynamics of each oscillator, $L = (l_{ij})_{i,j=1,\ldots,N}$ is the graph Laplacian of the underlying graph, and $u_i(t)$, $i = 1, \ldots, N$ are the controls. For the system to be well defined, we have to specify some initial conditions $x_i(t_0) = x_{i0}$ for $i = 1, \ldots, N$.

Assume that we have a reference trajectory $s(t)$, $t \geq t_0$ satisfying the differential equation

$$s'(t) \quad = \quad f(s(t)).$$

We want to control the system using a limited number of nodes. The selected nodes are called the "pinned nodes". For this purpose, we use a linear feedback law of the form

$$u_i(t) \quad = \quad p_i K e_i(t),$$

where $e_i(t) = s(t) - x_i(t)$, K is a feedback gain matrix, and where

$$p_i \quad = \quad \begin{cases} 1 & \text{if node } i \text{ is pinned} \\ 0 & \text{otherwise.} \end{cases}$$

Let P denote the diagonal matrix with diagonal vector $p_1, \ldots, p_N$.

The authors then give the definition of (global pinning-) controllability (based on Lyapunov stability criteria):

Definition 5.1 *We say that the system (5.-23) is controllable if the error dynamical system $e := (e_i(t))_{1 \leq i \leq N}$ is Lyapunov stable around the origin, i.e. there exists a positive definite function V such that $\frac{d}{dt} V(e(t)) < 0$ when $e(0) \neq 0$.*

The following result, [31, Corollary 5], provide a sufficient condition for a system to be controllable:

Proposition 5.2 ([31]) *Assume that f is such that there exists a bounded matrix $F_{\xi,\widetilde{\xi}}$, whose coefficients depend on ξ and $\tilde{\xi}$, which satisfies*

$$F_{\xi,\widetilde{\xi}}\left(\xi - \widetilde{\xi}\right) \quad = \quad f(\xi) - f(\widetilde{\xi}), \quad \xi, \widetilde{\xi} \in \mathbb{R}^n. \tag{5.-25}$$

Let $Q \in \mathbb{R}^{n \times n}$ be a positive definite matrix such that

$$\begin{array}{rcl} QK + K^t Q^t & = & \kappa \left(QB + B^t Q^t\right) \\ \left(QB + B^t Q^t\right) & \succeq & 0 \end{array}$$

and

$$\frac{1}{2}\,\lambda_N\left(\sigma L+\kappa P\right)\;\lambda_n\left(QB+B^tQ^t\right) \;>\; \sup_{\xi,\widetilde{\xi}}\|F_{\xi,\widetilde{\xi}}\|\;\|Q\|. \tag{5.-26}$$

Then the system is controllable.

Many systems of interest satisfy the constraint specified by (5.-25); see [22]. This proposition is very useful for node selection via the matrix P. Indeed, assume that Q is selected, then one may try to maximise $\lambda_N\left(\sigma L+\kappa P\right)$ as a function of P, under the constraint that no more than r nodes can be pinned. This is a combinatorial problem that can be relaxed using semi-definite programming or various heuristics [18].

Using Theorem 3.1, we are in position for stating an easy controllability condition in the spirit of [31, Corollary 7], based on the algebraic connectivity of the graph, the number of pinned nodes, the coupling strengh and the feedback gain.

Proposition 5.3 *If some positive definite symetric matrix $Q\in\mathbb{R}^{n\times n}$ is given that satisfies*

$$\begin{aligned} QK+K^tQ^t &= \kappa\left(QB+B^tQ^t\right)\\ \left(QB+B^tQ^t\right) &\succeq 0 \end{aligned}$$

and if κ satisfies

$$\kappa \;\geq\; \frac{\sum_{i=1}^{r} deg_i}{\sigma\lambda_{\min>0}(L)-\frac{2\;\|F_{\xi,\widetilde{\xi}}\|\;\|Q\|}{\lambda_{\min}(QB+B^tQ^t)}}+\sigma\lambda_{\min>0}(L),$$

then the system is controllable.

extbfProof. We follow the same steps as for the proof of Corollary 7 in [31]. We assume without loss of generality that the first r nodes are the pinned nodes. We may write P as

$$P \;=\; \sum_{i=1}^{r} e_ie_i^t,$$

where e_i is the i^{th} member of the canonical basis of $\mathbb{R}^N$, i.e. $e_i(j)=\delta_{i,j}$. We will try to compare $\lambda_N\left(\sigma L+\kappa P\right)$ with $\lambda_N\left(\sigma L\right)$ and use Proposition 5.2 to obtain a sufficient condition for controllability based on L, i.e. the topology of the network. For this purpose, let us notice recall that L can be written as

$$L \;=\; \mathcal{I}\cdot\mathcal{I}^t,$$

where $\mathcal{I}$ is the incidence matrix of any directed graph obtained from the system's graph by assigning an arbitrary sign to the edges [8]. Of course L will not depend on the chosen assignment. Using this factorization of L, we obtain that

$$\sigma L+\kappa\sum_{i=1}^{r}e_ie_i^t \;=\; \left[\sqrt{\kappa}\;e_r,\ldots,\sqrt{\kappa}\;e_1,\sqrt{\sigma}\mathcal{I}\right]\;\left[\sqrt{\kappa}\;e_r,\ldots,\sqrt{\kappa}\;e_1,\sqrt{\sigma}\mathcal{I}\right]^t.$$

Moreover, $\lambda_{\min>0}\left(\sigma L+\kappa P\right)$ can be expressed easily as the smallest nonzero eigenvalue of the r^{th} term of a sequence of matrices with shape (1.1) for with we can use Theorem 3.4 iteratively. Indeed, we have

$$\lambda_{\min>0}\left(\sigma L+\kappa e_1\right) \;=\; \lambda_{\min>0}\left(\left[\sqrt{\kappa}\;e_1,\sqrt{\sigma}\mathcal{I}\right]^t\left[\sqrt{\kappa}\;e_1,\sqrt{\sigma}\mathcal{I}\right]\right).$$

Let us denote by x the vector $\sqrt{\kappa}\ e_1$ and by X the matrix $[\sqrt{\sigma}\mathcal{I}]$. Then, we have that

$$\left[\sqrt{\kappa}\ e_1, \sqrt{\sigma}\mathcal{I}\right]^t \left[\sqrt{\kappa}\ e_1, \sqrt{\sigma}\mathcal{I}\right] \quad = \quad \begin{bmatrix} x^t x & x^t X \\ X^t x & X^t X \end{bmatrix}.$$

Therefore, Theorem 3.4 gives

$$\lambda_{\min>0}\left(\sigma L + \kappa e_1 e_1^t\right) \quad \geq \quad \sigma\lambda_{\min>0}(L) - \frac{deg_1}{(\kappa - \sigma\lambda_{\min>0}(L))},$$

where deg_1 is the degree of node number 1.

Let us now consider $\lambda_{\min>0}\left(\sigma L + \kappa\ e_1 + \delta_2 e_2\right)$. We have that

$$\lambda_{\min>0}\left(\sigma L + \kappa\ e_1 + \delta_2 e_2\right) \quad = \quad \lambda_{\min>0}\left(\left[\sqrt{\kappa}\ e_2, \sqrt{\kappa}\ e_1, \sqrt{\sigma}\mathcal{I}\right]^t \left[\sqrt{\kappa}\ e_2, \sqrt{\kappa}\ e_1, \sqrt{\sigma}\mathcal{I}\right]\right).$$

Let us denote by x the vector $\sqrt{\kappa}\ e_2$ and by X the matrix $[\sqrt{\kappa}\ e_1, \sqrt{\sigma}\mathcal{I}]$. Then, we have that

$$\left[\sqrt{\kappa}\ e_2, \sqrt{\kappa}\ e_1, \sqrt{\sigma}\mathcal{I}\right]^t \left[\sqrt{\kappa}\ e_2, \sqrt{\kappa}\ e_1, \sqrt{\sigma}\mathcal{I}\right] \quad = \quad \begin{bmatrix} x^t x & x^t X \\ X^t x & X^t X \end{bmatrix}$$

and using Theorem 3.4 again, we obtain

$$\lambda_{\min>0}\left(\sigma L + \kappa e_1 e_1^t + \kappa e_2 e_2^t\right) \quad \geq \quad \lambda_{\min>0}(\sigma L + \kappa e_1 e_1^t) - \frac{deg_2}{(\kappa - \lambda_{\min>0}(\sigma L + \kappa e_1 e_1^t))}.$$

Since $\lambda_{\min>0}(\sigma L + \kappa e_1 e_1^t) \leq \lambda_{\min>0}(\sigma L)$, we thus obtain

$$\lambda_{\min>0}\left(\sigma L + \kappa e_1 e_1^t + \kappa e_2 e_2^t\right) \quad \geq \quad \lambda_{\min>0}(\sigma L + \kappa e_1 e_1^t) - \frac{deg_2}{(\kappa - \sigma\lambda_{\min>0}(L))}.$$

We can repeat the same argument r times and obtain

$$\lambda_{\min>0}\left(\sigma L + \kappa P\right) \quad \geq \quad \sigma\lambda_{\min>0}(L) - \frac{\sum_{i=1}^r deg_i}{\kappa - \sigma\lambda_{\min>0}(L)}. \tag{5.-38}$$

Finally, by Proposition 5.2, we know that the following constraint is sufficient for preserving controllability

$$\lambda_{\min>0}\left(\sigma L + \kappa \sum_{i=1}^r e_i e_i^t\right) \quad \geq \quad \frac{2\ \|F_{\xi,\tilde{\xi}}\|\ \|Q\|}{\lambda_{\min}\left(QB + B^t Q^t\right)}. \tag{5.-37}$$

By (5.-38), it is sufficient to garantee the controllability of our system to impose

$$\sigma\lambda_{\min>0}(L) - \frac{\sum_{i=1}^r deg_i}{\kappa - \sigma\lambda_{\min>0}(L)} \quad \geq \quad \frac{2\ \|F_{\xi,\tilde{\xi}}\|\ \|Q\|}{\lambda_{\min}\left(QB + B^t Q^t\right)}.$$

This last inequality can then be written as

$$\kappa \quad \geq \quad \frac{\sum_{i=1}^r deg_i}{\sigma\lambda_{\min>0}(L) - \frac{2\ \|F_{\xi,\tilde{\xi}}\|\ \|Q\|}{\lambda_{\min}(QB + B^t Q^t)}} + \sigma\lambda_{\min>0}(L).$$

□

Bibliography

[1] Alon, N., Milman, V.D., λ_1, Isoperimetric inequalities for graphs, and superconcentrators, Journal of Combinatorial Theory, Series B, Volume 38, Issue 1, February 1985, Pages 73-88. (p. 302).

[2] Batson, Joshua D.; Spielman, Daniel A.; Srivastava, Ni. Twice-Ramanujan sparsifiers. STOC'09—Proceedings of the 2009 ACM International Symposium on Theory of Computing, 255–262, ACM, New York, 2009. (pp. 38 et 291).

[3] Benaych-Georges, Florent; Nadakuditi, Raj Rao. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. Adv. Math. 227 (2011), no. 1, 494–521. (pp. 38 et 291).

[4] Blank, Luise, Sarbu, Lavinia and Stoll, Martin, Preconditioning for Allen-Cahn variational inequalities with non-local constraints. J. Comput. Phys. 231 (2012), no. 16, 5406–5420. (pp. 38, 291 et 293).

[5] Boucheron, S., Lugosi, G., and Massart, P., Concentration inequalities using the entropy method. Ann. Probab. 31 (2003), no. 3, 1583–1614.

[6] S. Boucheron, G. Lugosi, and P. Massart, Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, 2013.

[7] Brémaud, P., Markov chains. Gibbs fields, Monte Carlo simulation, and queues. Texts in Applied Mathematics, 31. Springer-Verlag, New York, 1999.

[8] Brouwer, Andries E. and Haemers, Willem H. Spectra of graphs. Universitext. Springer, New York, 2012.

[9] Candès, E.J., The restricted isometry property and its implications for compressed sensing. C. R. Math. Acad. Sci. Paris 346 (2008), no. 9-10, 589–592. (p. 301).

[10] Candès, E. J. and Tao, T., Decoding by linear programming. IEEE Trans. Inform. Theory 51 (2005), no. 12, 4203–4215.

[11] Candès, Emmanuel J.; Romberg, Justin K.; Tao, Terence Stable signal recovery from incomplete and inaccurate measurements. Comm. Pure Appl. Math. 59 (2006), no. 8, 1207–1223.

[12] Candès, Emmanuel J.; Tao, Terence Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inform. Theory 52 (2006), no. 12, 5406–5425.

[13] Candès, E., Romberg, J. and Tao T., Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Information Theory, 2006, 2, 52, pp. 489–509. (p. 301).

[14] Candès, E. J. and Plan, Y. Near-ideal model selection by ℓ_1 minimization. Ann. Statist. 37 (2009), no. 5A, 285–2177.

[15] Chiang, Chen-Fu; Gomez, Guillermo Hitting time of quantum walks with perturbation. Quantum Inf. Process. 12 (2013), no. 1, 217–228.

[16] Chrétien, Stéphane; Corset, Franck. Least squares reconstruction of binary images using eigenvalue optimization. Signal Processing, 89 (2009) no. 11, 2079–2091. (p. 302).

[17] Dickson, K. I., Kelley, C. T., Ipsen, I. C. F. and Kevrekidis, I. G., Condition estimates for pseudo-arclength continuation. SIAM J. Numer. Anal. 45 (2007), no. 1, 263–276. (pp. 38, 291 et 304).

[18] Ghosh, A., and Boyd, S., Growing well connected graphs, Proceedings of the 45^{th} IEEE Conference on Decision and Control, (2006), p. 6605–6611. (p. 291).

(p. 300).

[19] Hoory, S., Linial, N. and Wigderson, A., Expander graphs and their applications. Bull. Amer. Math. Soc. (N.S.) 43 (2006), no. 4, 439–561. (p. 300).

[20] Horn, Roger A.; Johnson, Charles R. Matrix analysis. Cambridge University Press, Cambridge, 1985. (pp. 250 et 300).

[21] Ipsen, I. C. F. and Nadler, B. Refined perturbation bounds for eigenvalues of Hermitian and non-Hermitian matrices. SIAM J. Matrix Anal. Appl. 31 (2009), no. 1, 40–53. (p. 300).

[22] Jiang, G.-P., Tang, W. K.-S. and Chen, G., A simple global synchronization criterion for coupled chaotic systems. Chaos Solitons Fractals 15 (2003), no. 5, 925–935. (pp. 177, 178, 180, 181, 182, 183, 184, 185, 187, 188, 189, 192, 193, 205, 206, 207, 217, 218, 219, 220, 228, 229, 230, 233, 243, 245, 250, 254, 255, 256, 286, 300 et 301).

[23] Koltchinskii, V. and Mendelson, S., Bounding the smallest singular value of a random matrix without concentration, arXiv:1312.3580. (p. 293).

[24] Li, Chi-Kwong; Li, Ren-Cang. A note on eigenvalues of perturbed Hermitian matrices. Linear Algebra Appl. 395 (2005), 183–190. (p. 295).

[25] U. von Luxburg, A tutorial on spectral clustering, Stat. Comp. Vol. 17, (2007), no.4, 395–416.

[26] Merris, R., A note on Laplacian graph eigenvalues. Linear Algebra Appl. 285 (1998), no. 1-3, 33–35. (p. 291).

[27] Mohar, Bojan, Laplace eigenvalues of graphs–a survey Discrete Mathematics, Volume 109, Issues 1–3, 12 November 1992, Pages 171–183.

[28] Nadler, B., Finite sample approximation results for principal component analysis: a matrix perturbation approach. Ann. Statist. 36 (2008), no. 6, 2791–2817.

[29] Newman, M. E. J., Networks. An introduction. Oxford University Press, Oxford, 2010.

[30] Paouris, G., Concentration of mass on convex bodies, Geom. Funct. Anal. 16 (2006), 1021–1049.

[31] Porfiri, Maurizio and di Bernardo, Mario, Criteria for global pinning-controllability of complex networks. Automatica J. IFAC 44 (2008), no. 12, 3100–3106.

[32] Qiu, H., Pham, D.-S., Venkatesh, S. Lai, J. and Liu, W., Innovative sparse representation algorithms for robust face recognition, I8nternational Journal of Innovative Computing, Information and Control, 7 (2011), 10, p. 5645–5667.

[33] Rossberg, A.G., Food Webs and Biodiversity: Foundations, Models, Data, Wiley, 2013. (p. 304).

(p. 302).

[34] Shen, L. and Suter, B.W., Bounds for eigenvalues of arrowhead matrices and their applications to hub matrices and wireless communications, EURASIP Journal on Advances in Signal Processing 2009, 58. (p. 291).

[35] Tao, T., Notes 3a: Eigenvalues and sums of Hermitian matrices, http://terrytao.wordpress.com/2010/01/12/254a-notes-3a-eigenvalues-and-sums-of-hermitian-matrices

[36] Tropp, J. A. On the conditioning of random subdictionaries, Applied and Computational Harmonic Analysis, 25 (2008), no. 1, 1–24.

[37] Tropp, J. A. Norms of random submatrices and sparse approximation. C. R. Math. Acad. Sci. Paris 346 (2008), no. 23-24, 1271–1274. (pp. 39 et 293).

[38] Zhang W., Abreu, G., Inamori, M., Sanada, Y., Spectrum Sensing Algorithms via Finite Random Matrices, IEEE Trans. Communications, 60 (2012), no. 1, p. 164–175. (p. 304).

Chapter M

On the spacings between the successive zeros of the Laguerre polynomials

(pp. 39, 290, 293 et 294).
(p. 302).

with Sébastien Darses.

(p. 302).

Abstract

We propose a simple uniform lower bound on the spacings between the successive zeros of the Laguerre polynomials $L_n^{(\alpha)}$ for all $\alpha > -1$. Our bound is sharp regarding the order of dependency on n and α in various ranges. In particular, we recover the orders given in [1] for $\alpha \in (-1, 1]$.

(pp. 38, 39, 291 et 292).
(p. 302).

(pp. 38, 291, 293, 303 et 304).

1 Introduction

(p. 291).

The study of orthogonal polynomials has a long history with exciting interplay with numerous fields, including random matrix theory. The Laguerre polynomials which occur as the solutions of important differential equations [13], have had many applications in physics (electrostatics, quantum mechanics [6]), engineering (control theory; see e.g. [2]), random matrix theory (Wishart distribution; see e.g. [3] and [5]) and many other fields. The knowledge of the spacings between successive zeros of the Laguerre polynomials, interesting in its own right, is also potentially of great interest in many situations, e.g. for the spacings between successive eigenvalues of Wishart matrices, for bounding the gaps between sucessive energy levels in quantum mechanics or for the analysis of numerical algorithms in system identification problems, to name a few. (p. 302).

In this short note, we provide a uniform lower bound for the gaps between successive zeros of the Laguerre polynomials $L_n^{(\alpha)}$. In [1], important bounds were proposed in the case $\alpha \in (-1, 1]$ for individual spacings (i.e. bounds depending also on the ranking). Our bound is uniform but it is valid on the entire range $\alpha > -1$. For this reason, our bound might be helpful in a large number of applications. In particular, the cases including large values of α, are those of interest for random matrices with Wishart distribution. Our approach is based on a remarkable well known identity (a Bethe *ansatz* equation; see e.g. [11],[12]). (p. 291).

2 Preliminaries: Bethe ansatz equality

(pp. 38 et 292).

We first recall the following remarkable general result, see e.g. Lemma 1 in [11]. Let f be a polynomial with real simple zeros $x_1 < \cdots < x_n$, satisfying the ODE $f'' - 2af' + bf = 0$ where a and b are meromorphic function whose poles are different from the x_i's. Then for any fixed $k \in \{1 \cdots n\}$,

$$\sum_{j \neq k} \frac{1}{(x_k - x_j)^2} \quad = \quad \frac{\Delta(x_k) - 2a'(x_k)}{3}, \tag{2.1}$$

with $\Delta(x) = b(x) - a^2(x)$. Such equalities are called Bethe *ansatz* equations. (p. 300).

For $\alpha > -1$, the Laguerre polynomials $L_n^{(\alpha)}$ (n indicates the degree) are orthogonal polynomials with respect to the weight $x^\alpha e^{-x}$ on $(0, \infty)$. Let $x_{n,n}(\alpha) < \cdots < x_{n,1}(\alpha)$ denote the zeros of $L_n^{(\alpha)}$. It is known that the polynomial $L_n^{(\alpha)}$ is a solution of the second order ODE:

$$u'' - \left(1 - \frac{\alpha + 1}{x}\right) u' + \frac{n}{x} u \quad = \quad 0.$$

In this case, $a(x) = \frac{1}{2}\left(1 - \frac{\alpha+1}{x}\right)$. Therefore,

$$\Delta(x) = \frac{n}{x} - \frac{(x - \alpha - 1)^2}{4x^2} = \frac{-x^2 + (2(\alpha + 1) + 4n)x - (\alpha + 1)^2}{4x^2},$$

and then using the notations in [11],

$$\Delta(x) \quad = \quad \frac{(U^2 - x)(x - V^2)}{4x^2}, \tag{2.0}$$

where

$$U = \sqrt{n + \alpha + 1} + \sqrt{n}, \quad V = \sqrt{n + \alpha + 1} - \sqrt{n}. \tag{2.0}$$

(pp. 181, 187, 193, 218, 219, 222, 282 et 300).

Since the l.h.s. of (2.1) is positive and $a'(x) > 0$ for $x > 0$, an immediate consequence of (2.1) is that for all k, $(U^2 - x_{n,k}(\alpha))(x_{n,k}(\alpha) - V^2) > 0$, i.e.

$$V^2 < x_{n,n}(\alpha) < x_{n,1}(\alpha) < U^2. \tag{2.1}$$

Several bounds for the extreme zeros are known and can be found in [4, 7, 10, 11, 13]. For instance, using the Bethe *ansatz*, Krasikov proved [11, Theorem 1]:

$$V^2 + 3V^{4/3}(U^2 - V^2)^{-1/3} \leq x_{n,n}(\alpha) < x_{n,1}(\alpha) \leq U^2 - 3U^{4/3}(U^2 - V^2)^{-1/3} + 2. \tag{2.1}$$

(pp. 291 et 293).

3 Main result

We show by means of elementary computations that the Bethe *ansatz* equality actually yields a simple uniform lower bound for $x_{n,k}(\alpha) - x_{n,k+1}(\alpha)$, which turns out to be sharp, see Remark (2) below.

Theorem 3.1 *Let $\alpha > -1$. Then, the following lower bound for the spacings holds for all $k \in \{1, \cdots, n-1\}$:*

$$x_{n,k}(\alpha) - x_{n,k+1}(\alpha) \quad \geq \quad \sqrt{3}\frac{\alpha+1}{\sqrt{n(n+\alpha+1)}}. \tag{3.2}$$

Moreover, if $\alpha \geq n/C$ for some $C > 0$, we have

$$x_{n,k}(\alpha) - x_{n,k+1}(\alpha) \quad \geq \quad \frac{1}{\sqrt{C+1}}\sqrt{\frac{\alpha}{n}}. \tag{3.3}$$

Proof of Theorem 3.1

From (2.1), (2.1) and $a'(x) > 0$ for $x > 0$, we deduce the following inequality

$$\frac{1}{(x_{n,k}(\alpha) - x_{n,k+1}(\alpha))^2} \leq \sum_{j \neq k} \frac{1}{(x_{n,k}(\alpha) - x_{n,j}(\alpha))^2} \leq \frac{1}{3} \sup_{V^2 \leq x \leq U^2} \Delta(x). \tag{3.4}$$

The first inequality above seems to be crude, but is not, see Remark (1) below.

Let us then study the function Δ. The derivative of Δ on $(0, +\infty)$ reads:

$$\Delta'(x) = \frac{(-2x + U^2 + V^2)x^2 - 2x\left(-x^2 + (U^2+V^2)x - U^2V^2\right)}{4x^4} = \frac{2U^2V^2 - (U^2+V^2)x}{4x^3}.$$

Thus, Δ has a unique maximum on $(0, +\infty)$ that is reached at $x^* = \frac{2U^2V^2}{U^2+V^2}$. We have:

$$\begin{aligned} U^2 - x^* &= \frac{U^4 - U^2V^2}{U^2+V^2} = U^2\frac{U^2-V^2}{U^2+V^2} \\ x^* - V^2 &= \frac{U^2V^2 - V^4}{U^2+V^2} = V^2\frac{U^2-V^2}{U^2+V^2}. \end{aligned}$$

Thus, we obtain by plugging into (2.0),

$$\sup_{V^2 \leq x \leq U^2} \Delta(x) = \Delta(x^*) = \frac{(U^2-V^2)^2}{16\ U^2V^2},$$

since one can check that $x^* \in (V^2, U^2)$. Moreover, from the expressions (2) of U and V:

$$\begin{aligned} U^2 - V^2 &= (U-V)(U+V) = 4\sqrt{n}\sqrt{n+\alpha+1} \\ UV &= \alpha + 1. \end{aligned}$$

Hence, plugging these last equalities in (3.4), we can write

$$\frac{1}{(x_{n,k}(\alpha) - x_{n,k+1}(\alpha))^2} \quad \leq \quad \frac{1}{3}\ \frac{4^2\ n(n+\alpha+1)}{16\ (\alpha+1)^2},$$

and finally

$$x_{n,k}(\alpha) - x_{n,k+1}(\alpha) \quad \geq \quad \sqrt{3}\frac{\alpha+1}{\sqrt{n(n+\alpha+1)}}.$$

Now assume that $n \leq C\alpha$. Then $n+\alpha+1 \leq (C+1)\alpha+1$. Therefore $\sqrt{n+\alpha+1} \leq \sqrt{2(C+1)\alpha}$, where we used $1 \leq C\alpha \leq (C+1)\alpha$. Hence

$$x_{n,k}(\alpha) - x_{n,k+1}(\alpha) \quad \geq \quad \sqrt{\frac{3}{2(C+1)}}\sqrt{\frac{\alpha}{n}},$$

which completes the proof of Theorem 3.1.

Remarks

(*a*) Notice that replacing the sum $\sum_{j\neq k}(x_{n,k}(\alpha)-x_{n,j}(\alpha))^{-2}$ by the single term $(x_{n,k}(\alpha)-x_{n,k+1}(\alpha))^{-2}$ does not deteriorate a priori the order of dependency on n and α of a uniform bound in k of $x_{n,k}(\alpha) - x_{n,k+1}(\alpha)$. Indeed, let $0 < \delta < x_{n,k}(\alpha) - x_{n,k+1}(\alpha)$ for all k, we have the following simple inequality for any fixed k:

$$\frac{1}{(x_{n,k}(\alpha) - x_{n,k+1}(\alpha))^2} \leq \sum_{j\neq k}\frac{1}{(x_{n,k}(\alpha) - x_{n,j}(\alpha))^2} \leq \sum_{j\neq k}\frac{1}{(\delta|j-k|)^2} \leq 2\frac{\pi^2}{6}\frac{1}{\delta^2}.$$

(*b*) Let us verify that our bound is sharp regarding the order of dependency on n and α in various ranges.
Case $\alpha \in (-1,1]$: Theorem 5.1 in [1] says that for all $\alpha \in (-1,1]$:

$$(n+(\alpha+1)/2)\,(x_{n,k}(\alpha) - x_{n,k+1}(\alpha)) \quad \xrightarrow{n\to\infty} \quad j^2_{\alpha,k+1} - j^2_{\alpha,k},$$

where $j_{\alpha,k}$ is the k-th zeros of the Bessel function $J_\alpha(x)$. But, for all $k \geq 1$, the following holds (see [8, Theorem 3] and [9, p.2]):

$$\pi \leq j_{\alpha,k+1} - j_{\alpha,k} \leq 2\pi$$
$$j_{\alpha,k+1} + j_{\alpha,k} \geq 2\sqrt{(k-1/4)^2\pi + \alpha^2} \geq 1+\alpha.$$

As a consequence, for small k, $x_{n,k}(\alpha) - x_{n,k+1}(\alpha) \sim C(\alpha)/n$, which is consistent with our bound (3.2).
Case $n \leq C\alpha$ for an absolute constant $C > 0$: Summing (3.3) over k yields

$$\begin{aligned}\frac{\sqrt{n\alpha}}{\sqrt{C+1}} \quad &\leq \quad \sum_{1\leq k\leq n-1}(x_{n,k}(\alpha) - x_{n,k+1}(\alpha)) = x_{n,1}(\alpha) - x_{n,n}(\alpha)\\ &\leq \quad U^2 - V^2 = 4\sqrt{n}\sqrt{n+\alpha+1} \leq 6\sqrt{C+1}\sqrt{n\alpha},\end{aligned}$$

which means that the bound (3.3) is sharp with respect to the orders of n and α up to a multiplicative constant.

Notice moreover that in full generality, C can be taken as a function of n with absolutely no change in the proof.

(*c*) Finally, since the Bethe *ansatz* equation (2.1) is a general equality for polynomials f with real simple zeros, satisfying the ODE $f'' - 2af' + bf = 0$, good prior bounds on the extreme zeros for such polynomials could be used to obtain similar results as Theorem 3.1.

4 Numerical results

We now provide numerical results on the successive spacings of the Laguerre polynomials $L_n^{(\alpha)}$ for various values of n and α.

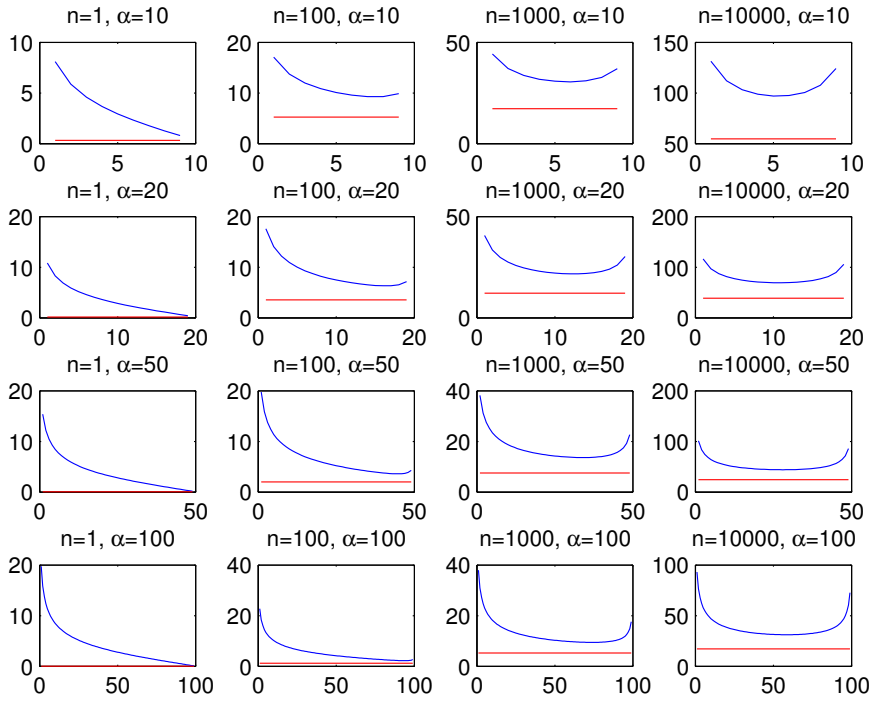

Figure M.1: Comparison between the uniform bound (3.2) in red, and the function $i \mapsto x_{n,i}(\alpha) - x_{n,i+1}(\alpha)$, $1 \leq i \leq n-1$, in blue. We set $\alpha = 1, 100, 10^3, 10^4$ and $n = 10, 20, 50, 100$.

Let us make a few comments on Figure M.1. The first column illustrates that the uniform bound almost coincides with the smallest spacing, which is here $x_{n,n-1}(1) - x_{n,n}(1)$ (Recall that $x_{n,n}(\alpha)$ is the smallest zero). When α is large compared to n, the behavior is quite different. For instance, based on Remark 2 in the case $\alpha \geq n/C$, we can expect most spacings to be almost equal, i.e. close to the uniform lower bound $\sqrt{\alpha/n}$ up to a multiplicative constant. In the last two columns of Figure M.1 (large values of α compared to n), we observe that this phenomena actually occurs in the bulk, i.e. for $\varepsilon n \leq i \leq (1-\varepsilon)n$, $0 < \varepsilon < 1$.

The results plotted in Figure M.1 have been obtained using Matlab and the codes available at http://people.sc.fsu.edu/~jburkardt/m_src/laguerre_polynomial/.

Bibliography

[1] Ahmed, S., Laforgia, A. and Muldoon, M. E. On the spacing of the zeros of some classical orthogonal polynomials. J. London Math. Soc. (2) 25 (1982), no. 2, 246–252. (pp. 309, 310 et 312).

[2] Datta, Kanti B. and Mohan, B. M. Orthogonal functions in systems and control. Advanced Series in Electrical and Computer Engineering, 9. World Scientific Publishing Co., Inc., River Edge, NJ, 1995. (p. 309).

[3] Dette, H. and Imhof, L., Uniform approximation of eigenvalues in Laguerre and Hermite β- ensembles by roots of orthogonal polynomials, Transactions of the AMS, 359 (2007), 10, 4999-5018. (p. 309).

[4] Dimitrov, D. K. and Nikolov, G. P Sharp bounds for the extreme zeros of classical orthogonal polynomials. J. Approx. Theory 162 (2010), no. 10, 1793–1804. (pp. 41 et 310).

[5] Faraut, J., Logarithmic Potential Theory, Orthogonal Polynomials, and Random Matrices CIMPA School, Hammamet, September 2011. Lecture notes available at http://www.math.jussieu.fr/ faraut/CIMPA-2011-JF.pdf. (p. 309).

[6] Freeden, Willi and Gutting, Martin, Special functions of mathematical (geo-)physics. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer Basel AG, Basel, 2013. (p. 309).

[7] Gatteschi, L. Asymptotics and bounds for the zeros of Laguerre polynomials: a survey. J. Comput. Appl. Math. 144 (2002), no. 1-2, 7–27. (pp. 41 et 310).

[8] Hethcote, H.W. Bounds for zeros of some special functions, Proc. Amer. Math. Soc. 25 (1970), 72-74. (p. 312).

[9] Finch, S. http://www.people.fas.harvard.edu/ sfinch/csolve/bs.pdf (p. 312).

[10] Ismail, Mourad E. H. and Li, X. Bound on the extreme zeros of orthogonal polynomials. Proc. Amer. Math. Soc. 115 (1992), no. 1, 131–140. (pp. 41 et 310).

[11] Krasikov, I. On extreme zeros of classical orthogonal polynomials. J. Comput. Appl. Math. 193 (2006), no. 1, 168–182. (pp. 41 et 310).

[12] Krasovsky, I. V., Asymptotic distribution of zeros of polynomials satisfying difference equations. J. Comput. Appl. Math. 150 (2003), no. 1, 56–70. (p. 310).

[13] Szego, G., Orthogonal polynomials, AMS (1975). (pp. 41, 309 et 310).

[14] R. Vershynin, Introduction to the non-asymptotic analysis of random matrices. Compressed sensing, 210–268, Cambridge Univ. Press, Cambridge, 2012.