(Non) convex losses and regularizations, some contributions in Statistics

Yohann de Castro

UNIVERSITÉ PARIS-SUD

Faculté des Sciences d'Orsay
École Doctorale de Mathématiques Hadamard (ED 574)
Laboratoire de Mathématiques d'Orsay (UMR 8628 CNRS)

Mémoire présenté pour l'obtention de

l'Habilitation à Diriger des Recherches

Discipline : Mathématiques

par

Yohann DE CASTRO

PERTES ET RÉGULARISATIONS (NON) CONVEXES, QUELQUES CONTRIBUTIONS EN STATISTIQUE.

Rapporteurs :
EMMANUEL CANDÈS
CHRISTOPHE GIRAUD
ÉRIC MOULINES
SARA VAN DE GEER

Date de soutenance : Mercredi 30 Novembre 2016

Composition du jury :

ÉLISABETH GASSIAT	(Examinatrice)
CHRISTOPHE GIRAUD	(Rapporteur)
GUILLAUME LECUÉ	(Examinateur)
VINCENT RIVOIRARD	(Examinateur)
JUDITH ROUSSEAU	(Examinatrice)
ALEXANDRE TSYBAKOV	(Examinateur)

Laboratoire de Mathématiques d'Orsay
Univ. Paris-Sud, CNRS, Université Paris-Saclay
91405 ORSAY CEDEX, FRANCE

Résumé

Ce mémoire couvre essentiellement les travaux menés par l'auteur depuis sa nomination en tant que "Maître de Conférences" au Laboratoire de Mathématiques d'Orsay (LMO), c'est-à-dire depuis la fin de son doctorat en décembre 2011 à l'Institut de Mathématiques de Toulouse (IMT). Durant cette période, l'auteur a renforcé ses contributions à la statistique en grandes dimensions et exploré de nouveaux champs de recherche qui peuvent être résumés autour des thématiques de la "super résolution" (ou plus généralement du problème extrémal des moments généralisés) et des "modèles à espace latent" (et en particulier les modèles des chaînes de Markov cachées). Ce manuscript ne cherche pas à présenter de manière exhaustive les résultats développés par l'auteur mais plutôt un point de vue synthétique sur ces contributions. La/le lectrice/lecteur est invité-e à consulté les articles cités pour plus de détail et un traitement mathématique plus précis des sujets présentés ici.

Mots clefs : Statistique en grandes dimensions ; Problème des moments ; Modèles de chaînes de Markov cachées ; Test d'hypothèses ; Sélection de Modèles ; Analyse de la sensibilité ; Optimisation convexe ; Matrices aléatoires ; Processus gaussiens ; Méthodes spectrales.

Abstract

This dissertation essentially covers the work done by the author as "Maître de Conférences" at the Laboratoire de Mathématiques d'Orsay (LMO), that is to say since the end of his Ph.D. thesis in December 2011 at the Institut de Mathématiques de Toulouse (IMT). During this period, the author strengthened his contributions to high-dimensional statistics and investigated new fields of research that may be summarized under the following labels "Super Resolution" (and more generally the extremal moment problem) and "latent space models" (and in particular the hidden Markov models). This report is not meant to present comprehensive description of the results developed by the author, but rather a synthetic view of his main contributions. The interested reader may consult the cited articles for further details and the precise mathematical treatment of the topics presented here.

Keywords: High-Dimensional Statistics; Moment problem; Hidden Markov models; Hypothesis testing; Model selection; Sensitivity analysis; Convex optimization; Random matrix; Gaussian process; Spectral methods.

Table des matières

Table des figures

Productions scientifiques

Les articles **[DC1]**, **[DC2]**, **[DC3]** et **[DC13]** correspondent aux travaux doctoraux de l'auteur menés entre 2009 et 2012 à l'Institut de Mathématiques de Toulouse. Les autres articles sont le fruit de ses recherches au Laboratoire de Mathématiques d'Orsay entre 2012 et 2016.

Articles dans des revues internationales à comité de lecture

[DC1] Yohann De Castro and Fabrice Gamboa. *Exact reconstruction using Beurling minimal extrapolation*. J. Math. Anal. Appl. 395 (2012), no. 1, 336–354.

[DC2] Yohann De Castro. *A remark on the lasso and the Dantzig selector.* Statist. Probab. Lett. 83 (2013), no. 1, 304–314.

[DC3] Yohann De Castro. *Optimal designs for lasso and Dantzig selector using expander codes*. IEEE Trans. Inform. Theory 60 (2014), no. 11, 7293–7299.

[DC4] Flavia Barsotti, Yohann De Castro, Thibault Espinasse, and Paul Rochet. *Estimating the transition matrix of a Markov chain observed at random times.* Statist. Probab. Lett. 94 (2014), 98–105.

[DC5] Jean-Marc Azaïs, Yohann De Castro, and Fabrice Gamboa. *Spike detection from inaccurate samplings.* Appl. Comput. Harmon. Anal. 38 (2015), no. 2, 177–195.

[DC6] Yohann De Castro and Guillaume Mijoule. *Non-uniform spline recovery from small degree polynomial approximation.* J. Math. Anal. Appl. 430 (2015), no. 2, 971–992.

[DC7] Yohann De Castro and Alexandre Janon. *Randomized pick-freeze for sparse Sobol indices estimation in high dimension.* ESAIM Probab. Stat. 19 (2015), 725–745.

[DC8] Yohann De Castro, Élisabeth Gassiat, and Claire Lacour. *Minimax adaptive estimation of non-parametric hidden markov models.* J. Mach. Learn. Res., 17 (2016), no. 111, 1–43.

[DC9] Jean-Marc Azaïs, Yohann De Castro, and Stéphane Mourareau. *A Rice method proof of the Null-Space Property over the Grassmannian*. Ann. Inst. H. Poincaré Probab. Stat., to appear (2016).

[DC10] Jean-Marc Azaïs, Yohann De Castro, and Stéphane Mourareau. *Power of the Spacing test for Least-Angle Regression*. Bernoulli, to appear (2016).

[DC11] Yohann De Castro, Fabrice Gamboa, Didier Henrion, and Jean-Bernard Lasserre. *Exact solutions to Super Resolution on semi-algebraic domains in higher dimensions*. IEEE Trans. Inform. Theory, to appear (2016).

[DC12] Claire Boyer, Yohann De Castro and Joseph Salmon. *Adapting to unknown noise level in sparse deconvolution*. Information and Inference : a Journal of the IMA, to appear (2016).

Articles dans des revues nationales à comité de lecture

[DC13] Yohann De Castro. *Quantitative isoperimetric inequalities on the real line.* Ann. Math. Blaise Pascal 18 (2011), no. 2, 251–271.

Articles soumis à des revues internationales à comité de lecture

[DC14] Yohann De Castro, Élisabeth Gassiat, and Sylvain Le Corff. *Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models*. Submitted, (2016) (arXiv :1507.06510).

[DC15] Yohann De Castro, Thibault Espinasse, and Paul Rochet. *Reconstructing undirected graphs from eigenspaces*. Submitted, 2016 (arXiv :1603.08113).

[DC16] Sandrine Dallaporta and Yohann De Castro. *Restricted Isometry Constants for Gaussian and Rademacher matrices*. Submitted, 2016 (arXiv :1604.01171).

[DC17] Yohann De Castro, Yannig Goude, Georges Hébrail, Jiali Mei. *Recovering Multiple Time Series From a Few Temporally Aggregated Measurements*. Submitted, 2016.

Travaux en cours ou sur le point d'être soumis

[DC18] Yohann De Castro, Fabrice Gamboa, Didier Henrion, and Jean-Bernard Lasserre. *Approximate optimum designs on semi-algebraic design spaces*. In progress, 2016.

[DC19] Jean-Marc Azaïs, Yohann De Castro, and Stéphane Mourareau. *Spacing test for Gaussian processes and Super-Resolution*. In progress, 2016.

[DC20] Yohann De Castro, Claire Lacour and Thanh Mai Pham Ngoc. *Estimating Structured Graphons*. In progress, 2016.

Thèse de Doctorat

[Thèse] Yohann De Castro. *Constructions déterministes pour la régression parcimonieuse*. Thèse de Doctorat, Université Paul Sabatier, Toulouse. Décembre 2011.

Brevet

[Brevet] Jean-Marc Azaïs, Yohann De Castro, Yannig Goude, Georges Hébrail, Jiali Mei. *Procédé d'estimation de consommation/production de fluides/effluents à partir de mesures partielles*. 2016.

Préambule

Parcours scientifique

Les travaux présentés dans ce mémoire ont commencé par mes rencontres avec Yves Meyer (en 2007) puis Emmanuel Candès (en 2008) qui ont bien voulu encadrer mes mémoires de master sur un aperçu de la théorie du "*Compressed Sensing*" dont l'objet est la reconstruction de signaux en grandes dimensions à partir de quelques mesures linéaires aléatoires. L'enthousiasme d'Yves Meyer et la bienveillance d'Emmanuel Candès sont de précieux souvenirs de mes premiers regards sur la Statistique mathématique.

Par la suite, mes travaux doctoraux (2009-2012) m'ont amenés à l'Institut de Mathématiques de Toulouse (IMT) au contact de Jean-Marc Azaïs et Franck Barthe afin de travailler sur les "*constructions déterministes pour la régression parcimonieuse*" **[Thèse]**. La question principale de cette thèse était de proposer des plans d'expériences déterministes pouvant se substituer aux mesures aléatoires qui ont fait le succès du Compressed Sensing. Cette question redoutable m'a sensibilisé à des problématiques passionnantes comme le problème de la transition de phase en minimisation ℓ_1 **[DC9]**, les conditions d'isométries restreintes **[DC2]**, les codes correcteurs d'erreurs **[DC3]** ou encore la reconstruction de mesures signées à partir de moments généralisés **[DC1]** **[DC5]** sous l'impulsion de Fabrice Gamboa.

En septembre 2012, j'ai eu le plaisir de rejoindre le Laboratoire de Mathématiques d'Orsay (LMO) en tant que Maître de Conférences (2012-Présent). J'y ai découvert une équipe accueillante et stimulante. Au fil des groupes de travail sur les matrices aléatoires, j'ai pu collaborer avec Sandrine Dallaporta (École Normale Supérieure de Cachan) sur des problèmes de constantes d'isométrie pour des modèles de Wishart sous gaussiens **[DC16]**. De même, j'ai eu la chance d'élargir mon cadre de recherche grâce à Claire Lacour, Élisabeth Gassiat **[DC8]** et Sylvain Le Corff **[DC14]** en travaillant sur l'estimation non paramétrique des chaînes de Markov cachées. J'y ai aussi rencontré Alexandre Janon qui m'a initié à l'analyse de la sensibilité, à la méthode de Sobol et celle du "*Pick-Freeze*" **[DC7]**. Je terminerai ce paragraphe orcéen par une collaboration naissante **[DC20]** avec Claire Lacour et Thanh Mai Pham Ngoc sur l'estimation non paramétrique de graphes géométriques.

Parallèlement à mon arrivée à Orsay, j'ai commencé à travailler avec Didier Henrion et Jean-Bernard Lasserre du Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS, Toulouse), ainsi qu'avec Fabrice Gamboa, sur l'utilisation de "*hiérarchies*" pour des problèmes extrémaux de moments en Statistique **[DC11]**, voir pour la construction de plans d'expériences optimaux **[DC18]**. Avec Thibault Espinasse (Université Lyon 1) et Paul Rochet (Université de Nantes), nous travaillons sur la reconstruction de graphes à partir des vecteurs propres de leurs matrices d'adjacence **[DC15]** et la reconstruction de matrices de transition **[DC4]** avec Flavia Barsotti (UniCredit, Milan). Au cours des années de thèse de Stéphane Mourareau, j'ai eu le plaisir de collaborer avec lui et Jean-Marc Azaïs sur l'utilisation de la "*méthode de Rice*" en statistique en grandes dimensions **[DC9]** et les tests "post-inférence" **[DC10]**. Récemment, j'ai converti Guillaume Mijoule (Université de Liège) **[DC6]**, Claire Boyer (IMT, Toulouse) et Joseph Salmon (Telecom Paris) **[DC12]** au problème de la "*super résolution*".

Aperçu des travaux de recherche

Cette section résume mes principales contributions en essayant d'en dégager l'importance et l'originalité au vu de la bibliographie présentée dans ce mémoire. Je tiens à mettre en garde sur le fait que ce parti pris bibliographique ne pourrait se substituer à la lecture des articles incriminés que j'ai (co-)écrit, ni à celle des références que je pointe. Ce point de vue, forcément biaisé, essaie pour autant de présenter succinctement mes travaux de recherche de la manière la plus objective possible.

Problème extrémal des moments et super résolution

Dans l'article **[DC1]**, nous avons montré comment reconstruire exactement une mesure signée de support fini[1] à partir de la connaissance d'un nombre fini de moments généralisés (*i.e.,* moments algébriques, trigonométriques, coefficients de Fourier, évaluation de la transformée de Laplace ou de Stieltjes en quelques points, etc.) à l'aide d'un estimateur solution d'un programme de minimisation[2] ℓ_1, voir Section 2.1.2. En particulier, nous avons démontré que l'on pouvait avoir reconstruction exacte de la mesure cible si un certain problème d'interpolation polynomiale contraint pouvait être résolu, ce que nous avons fait dans les cas correspondant à la reconstruction exacte d'une mesure cible positive et celui d'une mesure signée quelconque où l'on observe un très grand nombre de moments généralisés[3]. Par la suite, dans un article incontournable [CFG14], Emmanuel Candès et Carlos Fernandez-Granda ont montré que, dans le cas d'observations données par des coefficients de Fourier[4], on pouvait résoudre ce problème d'interpolation (et donc avoir reconstruction exacte) à partir d'un nombre drastiquement[5] petit de coefficients de Fourier.

De plus, ils ont proposé une formulation semi-définie du programme de minimisation ℓ_1 dans [CFG13, CFG14], ce qui a permis d'ouvrir le champ des applications concrètes de cette théorie. Plus tard, on a présenté dans l'article **[DC11]**, l'utilisation de "*hiérarchies*" de Lasserre pour résoudre ce problème numériquement, voir Section 2.1.3. Cette nouvelle approche permet de certifier l'exactitude de la solution du programme de minimisation ℓ_1, là où les autres travaux ne peuvent pas être théoriquement utilisés. En effet, ceux-ci reposent sur l'utilisation du théorème de Riesz-Féjer et ne peuvent donc s'appliquer que pour des mesures dont le support est contenu sur $[0,1]^d$ où $d = 1,2$, alors que notre approche utilise un autre point de vue (celui du théorème de Putinar) qui permet de contourner ces limitations. Récemment, nous avons étendu **[DC18]** ces "*hiérarchies*" au cadre de construction de plans d'expériences optimaux (voir Section 2.3), ce qui ouvre un cadre de travail, nous semble-t-il, très prometteur.

Le problème important de l'estimation du support a été résolu simultanément par **[DC5]** et [FG13] avec l'utilisation clef d'un contrôle d'une divergence de Bregman dans **[DC5]**, qui dénote des techniques que l'on rencontre en statistique en grandes dimensions. La prédiction minimax a été obtenue par [TBR15] et étendue au cadre d'une variance inconnue par **[DC12]**. L'article **[DC12]** est le premier à s'intéresser à l'estimation simultanée[6] de la variance et de la mesure cible à l'aide des outils de preuve de la super résolution, voir Section 2.2.3. Pour être exhaustif, je pointerais l'article **[DC6]** qui est le premier à proposer une estimation "*sans grille*" des nœuds d'une approximation par splines avec conditions aux bords, à l'aide des techniques de "*super résolution*" présentées ici.

[1] *i.e.,* une somme finie de masses de Dirac avec poids réels.
[2] plus précisément, la norme en variation totale d'une mesure borélienne finie.
[3] à savoir un nombre exponentiel en l'inverse de la distance entre deux atomes de la mesure cible.
[4] *i.e.,* le problème de la "*super résolution*".
[5] ici, de l'ordre de l'inverse de la distance entre deux atomes de la mesure cible.
[6] connue en statistique en grandes dimension autour de l'estimateur "*Scaled-Lasso*" ou "*Square-root Lasso*".

Estimation et prédiction en grandes dimensions

Ce sujet très compétitif a connu un essor important au cours de la dernière décennie en Statistique. Les résultats sont basés sur une utilisation des conditions lagrangiennes d'optimalité (KKT) satisfaites par les estimateurs issus d'une minimisation convexe et des conditions sur la matrice de design du type "*Restricted Isometry Property*" (RIP). L'idée est alors de montrer que l'une de ces conditions sur le design de type RIP est satisfaite avec très grande probabilité pour des matrices de design aléatoires. Dans ma quête de plans déterministes, j'ai proposé la condition "*Universal Distortion Property*" (UDP) dans l'article **[DC2]**. Cette condition UDP est la moins restrictive[7] pour avoir des "*inégalités oracles*[8]" pour la perte ℓ_1 et la prédiction. Contrairement aux autres conditions, UDP ne fait pas intervenir de valeurs propres restreintes mais seulement la "*distortion*" du noyau et la plus petite valeur singulière non nulle de la matrice de design normalisée, voir Section 1.2.1. De plus **[DC3]**, lorsque le design est donné par la matrice d'adjacence normalisée d'un graphe expanseur déséquilibré[9], on peut lier UDP au facteur d'expansion du graphe.

Un autre phénomène passionnant en statistique en grandes dimensions décrit la "*transition de phase*" de la reconstruction exacte [DT09b, DT09a]. Il s'agit de préciser, en fonction des tailles du problème (parcimonie, nombre d'observations et taille du vecteur cible), quand la reconstruction exacte est possible (borne inférieure) ou impossible (borne supérieure). Étonnamment, ces deux bornes coïncident dans l'asymptotique pour des design gaussiens. Il existe très peu de preuves directes de la borne inférieure et nous en avons donné dans l'article **[DC9]** en utilisant la "*méthode de Rice*[10]" qui n'avait jamais été mise en œuvre en statistique en grandes dimensions, voir Section 1.1.2. Un enjeu similaire consiste à trouver des bornes précises sur les constantes intervenants dans la propriété RIP. À l'aide de la loi jointe des valeurs propres d'une matrice de Wishart gaussienne, cela a été fait dans [BCT11]. En utilisant les fonctions de taux d'inégalités de déviations classiques, nous avons obtenu **[DC16]** un résultat similaire et comparable dans un cadre plus général couvrant les matrices sous-gaussiennes, voir Section 1.2.2. Ce travail est, on l'espère, un préliminaire qui permettra d'accrocher un résultat d'"*universalité*" pour les constantes RIP.

Tests en grandes dimensions

De nouveaux tests sont apparus récemment en statistique en grandes dimensions, lire par exemple [vdG16, HTW15]. En particulier, les tests "*post-inférence*" permettent de tester conditionnellement au support et aux signes d'estimateurs construits par minimisation ℓ_1. Ces tests sont "*exacts*[11]" [TLTT14, LTTT14] et ont suscité un grand intérêt auprès de la communauté statistique. L'un des tests les plus simples à mettre en œuvre consiste à tester une hypothèse globale nulle en regardant l'écart entre les deux premiers nœuds du LARS, il s'agit du "*Spacing test*". Étonnamment, aucun calcul de puissance n'avait été fourni dans ce cadre et, *de facto*, aucune justification théorique de la zone de rejet proposée[12] par [TLTT14, LTTT14]. Dans l'article **[DC10]**, nous donnons une preuve élémentaire du "*Spacing test*" qui n'utilise pas la formule de Kac-Rice comme dans [TLTT14, LTTT14] ; nous prouvons que le test est non biaisé et que la zone de rejet est optimale[13] dans le cas d'un design orthonormé ; et nous démontrons que la "*studentisation*[14]" du "*Spacing test*" est possible et donne, sous l'hypothèse nulle, une loi de test uniforme "*exacte*" elle aussi, voir Section 1.3. Nous travaillons actuellement à étendre le "*Spacing test*" au cadre de la super résolution **[DC19]**.

[7] au sens où les autres conditions de la littérature l'impliquent.

[8] *i.e.*, certifier que l'estimateur *Lasso* et le *sélecteur de Dantzig* sont parmi les "meilleurs" estimateurs pour l'estimation et la prédiction en régression parcimonieuse.

[9] que l'on sait construire de manière déterministe [GUV09] en temps polynomial.

[10] utilisation d'une formule de Kac-Rice pour estimer la queue de distribution du maxima d'un processus.

[11] La loi de test sous l'hypothèse nulle est uniforme sur $[0,1]$ pour toutes tailles d'échantillon.

[12] La statistique de test est uniforme sur $[0,1]$ sous l'hypothèse nulle et la valeur observée est le niveau observé.

[13] au sens où elle donne la plus grande puissance pour toutes les alternatives.

[14] extension du test au cadre de la variance inconnue par une estimation indépendante de la variance.

Modèles de Markov cachés

L'identifiabilité des modèles de Markov cachés (HMM en anglais) a été démontrée récemment dans la cadre paramétrique [AMR09] et dans le cadre non paramétrique [GCR16, AH16]. Dans l'article **[DC8]**, nous proposons pour la première fois un estimateur *adaptatif minimax* pour l'estimation non paramétrique des lois d'émission en utilisant une méthode des moindres carrés pénalisés "*à la Birgé-Massart*". Pour cela, nous avons étendu l'estimateur spectral[15] [AHK12, HKZ12] au cadre non paramétrique. Ainsi, nous avons démontré que l'estimateur des moindres carrés pénalisés peut être efficacement utilisé dans la pratique en le combinant à l'estimateur spectral. Plus de détails sont donnés en Section 3.1.

Puis avec Élisabeth Gassiat et Sylvain Le Corff **[DC14]**, nous avons utilisé l'estimateur spectral pour l'estimation des états de la chaîne de Markov cachée, *i.e.*, l'estimation des lois a posteriori de "filtrage" et de "lissage", voir Section 3.1.5. Bien que remplacer les paramètres par leur estimation dans le calcul des lois a posteriori et l'inférence des états cachés est habituel dans la pratique, les résultats théoriques pour délimiter cet usage sont peu nombreux. Il nous semble que seul [EDKM07] étudie la distribution de filtrage dans un cadre paramétrique. L'article **[DC14]** comble ce manque en traîtant les cadres paramétrique et non paramétrique, ainsi que l'erreur d'estimation de la loi de filtrage et celle de lissage. Nous proposons, de plus, l'utilisation de la méthode spectrale qui, contrairement à l'algorithme "espérance-maximisation" (EM) par exemple, ne souffre pas de problème d'initialisation de l'algorithme. Nous en donnons l'erreur d'estimation des lois a posteriori dans un cadre non paramétrique.

Analyse de sensibilité

Dans l'article **[DC7]**, nous avons étendu l'approche "*Pick-Freeze*" de l'analyse de sensibilité à l'aide d'outils de statistique en grandes dimensions. Cette méthode permet d'estimer les "*indices de Sobol*[16]" des entrées d'une fonction sans hypothèse de régularité sur celle-ci. Pour l'étendre à des fonctions ayant potentiellement un grand nombre d'entrées mais dont peu sont "*influentes*", nous avons utilisé des outils de sélection de support à l'aide de l'estimateur Lasso seuillé, voir Section 1.2.3. Notre apport a été d'introduire le "*Randomized Pick-Freeze*", qui prédit simultanément plusieurs indices de Sobol choisis au hasard et donne, *in fine*, une estimation des entrées ayant un indice de Sobol significatif (*i.e.*, le support). Ce point de vue permet de se rattacher au problème d'estimation du support en grandes dimensions et, réciproquement, de motiver l'étude de certains designs[17] en Statistique.

Reconstruction de graphes

Dans l'article **[DC15]**, nous introduisons une nouvelle problématique : l'estimation d'un graphe à partir d'une observation erronée des vecteurs propres d'une matrice d'adjacence pondérée. À notre connaissance, cela n'a jamais été fait. Nous donnons une condition suffisante et une condition nécessaire pour que le modèle soit identifiable, et nous proposons une méthode *ad hoc* d'estimation basée sur l'étude théorique d'un nouveau critère, voir Section 3.2. Entre autres, ce travail couvre le cadre de **[DC4]** qui étudie l'estimation de la matrice de transition d'une chaîne de Markov observée à sauts de temps i.i.d. de loi inconnue.

Brevet sur l'estimation de la consommation d'électricité

Dans le cadre du contrat d'accompagnement CIFRE de Mme Mei (que je co-encadre) conclu entre ÉDF Saclay et l'Université Paris-Sud, nous avons déposé avec Jean-Marc Azaïs, Yannig Goude, Georges Hébrail et Jiali Mei, un brevet **[Brevet]** sur un "*procédé d'estimation de consommation/production de fluides/effluents à partir de mesures partielles*".

[15] qui connaît un succès tangible dans la communauté d'apprentissage statistique.

[16] sorte de mesure de l'influence d'une variable d'entrée sur les valeurs d'une fonction.

[17] les entrées de la matrice de design sont des Rademacher (ou des Bernoulli) liées lorsque l'on s'intéresse aux indices de Sobol des termes d'interactions entre les entrées de la fonction.

Plan du manuscrit

Ce manuscrit s'articule autour de trois chapitres regroupant les contributions respectives de l'auteur à la statistique en grandes dimension (chapitre 1), au problème extrémal des moments et à la super résolution (chapitre 2), et aux modèles de chaînes de Markov cachées et à la reconstruction de graphes à partir de vecteurs propres perturbés (chapitre 3).

Le chapitre 1 est basé sur les articles **[DC2]** **[DC3]** **[DC7]** **[DC9]** **[DC10]** et **[DC16]**. Il traite de l'étude par l'auteur de la propriété "Universal Distortion Property" (UDP) **[DC2]** **[DC3]**, la "null space property" (NSP) **[DC9]** et la "Restricted Isometry Property" (RIP) **[DC16]**. Il présente une nouvelle méthode en analyse de la sensibilité, le "Randomized Pick-Freeze", introduite par **[DC7]**. Enfin, il décrit les outils d'analyse de la puissance du "Spacing test" développés dans **[DC10]**.

Le chapitre 2 est écrit à partir de articles **[DC1]** **[DC5]** **[DC6]** **[DC11]** et **[DC12]**. En particulier, on y aborde la théorie de la "super résolution" à travers des résultat de reconstruction exacte **[DC1]**, de localisation **[DC5]** **[DC6]**, de prédiction "presque" minimax et estimation du niveau de bruit **[DC12]** et d'optimisation sur des domaines semi-algébriques **[DC11]**. Il s'appuie aussi sur un travail en cours **[DC18]** en décrivant une nouvelle méthodologie pour la construction de plans expérimentaux optimaux.

Le dernier chapitre est issu des articles **[DC4]** **[DC8]** **[DC14]** **[DC15]**. Il porte sur l'estimation adaptative minimax des chaînes de Markov cachées non paramétriques **[DC8]** et sur l'estimation des distributions conditionnelles des états de la chaîne cachée dans les cadres paramétrique et non paramétrique **[DC14]**. De manière indépendante, il aborde les travaux de l'auteur sur la reconstruction de graphes à partir de l'observation d'une perturbation des vecteurs propres d'une matrice d'adjacence pondérée **[DC4]** **[DC15]**.

Chapter 1

High-dimensional perspectives

A major development in modern Statistics has been brought by the idea that one can recover a high-dimensional target from a few linear observations by ℓ_1-minimization as soon as the target vector is "*sparse*" in a well-chosen basis. Undoubtedly, the notion of "*sparsity*" has encountered a large echo among the statistical community and many successful applications rely on ℓ_1-minimization—the reader may consult [CDS98, Tib96, Fuc05, CT06, CT07] for some seminal works, [HTFF05, BvdG11, CGLP12, FR13, Gir14] for a review and references therein. Some of the most popular estimators in high-dimensional Statistics remain the Lasso [Tib96] and the Dantzig selector [CT07]. Lot of interest has been dedicated to the estimation, prediction or support recovery problems using these estimators. This body of work has been developed around sufficient conditions on the design matrix such as *Restricted Isometry Property* [CT06], *Restricted Eigenvalue* [BRT09], *Compatibility* [vdGB09, BvdG11], *Universal Distortion* **[DC2]**, $\mathbf{H}_{s,1}$ [JN11], or *Irrepresentability* [Fuc05], to name but a few. Those conditions enclose the spectral properties of the design matrix on the set of (almost) sparse vectors. Using this spectral feature and exploiting the implicit optimality equation given by the Karush-Kuhn-Tucker (KKT) conditions, one can derive "*oracle inequalities*" and/or a control on the support recovery error.

Of course, this chapter is not devoted to an exhaustive presentation of high-dimensional Statistics but rather some recent points of interest that I have came across during my research in Orsay over the last four years. They fall into four parts: new proof of the null space property using the Kac-Rice Formula (Section 1.1.2), Sensitivity Analysis (Section 1.2.3), assessing oracle inequalities using deviations inequalities on Wishart matrices (Section 1.2.2), and hypothesis testing using exact Post-Selection Inference (Section 1.3).

1.1 Phase transition on exact recovery

1.1.1 The convex relaxation's breakthrough

One of the simplest inverse problems can be stated as follows. Given a matrix $X \in \mathbb{R}^{n\times p}$ and an observation $y \in \mathrm{Im}(X)$, can we faithfully recover β^0 such that $y = X\beta^0$ holds? In the ideal case where $n \geq p$ and the matrix X is one to one, this problem is elementary. However, in view of recent applications, the frame of high-dimensional Statistics is governed by the opposite situation where $n < p$. To bypass the limitations due to non identifiability, one usually assumes that the "*design*" matrix X is random [CRT06, CT06] and one considers [CDS98] the ℓ_1-minimization procedure

$$\mathbf{D}_X(\beta^0) \in \arg\min_{X\beta = X\beta^0} \|\beta\|_1 , \tag{1.1}$$

where $\beta^0 \in \mathbb{R}^p$ is a "*target*" vector we aim to recover.

The high-dimensional models often assume that the target vector β^0 is well approximated by the space of s-sparse vectors $\Sigma_s := \{\beta \in \mathbb{R}^p\,,\ \|\beta\|_0 \leq s\}$ where $\|\beta\|_0$ denotes the size of the support of β, *i.e.,* the set of nonzero coefficients. Note that this framework is the baseline of the

flourishing Compressed Sensing (CS), see [CRT06, Don06a, CDD09, CGLP12, FR13] and references therein. A breakthrough brought by CS states that if the matrix X is drawn at random (*e.g.,* X has i.i.d.standard Gaussian entries) then, with overwhelming probability, one can recover $\beta^0 \in \Sigma_s$ using (1.1). Precisely, the interplay between randomness and ℓ_1-minimization shows that with only n measurements such that

$$n \geq c_1 s \log\Big(\frac{c_2 p}{s}\Big), \tag{1.2}$$

where $c_1, c_2 > 0$ are numerical constants, one can faithfully reconstruct any s-sparse vector β^0 from the knowledge of X and an observation $y = X\beta^0$. Notably, this striking fact is governed by the null space property (NSP) defined as follows.

Definition (NSP(s, C))**.** *A space $G \subset \mathbb{R}^p$ satisfies the null space property of order s and dilatation $C \geq 1$ if and only if*

$$\forall\, h \in G\,, \quad \forall\, S \subset [p] \quad \text{s.t.} \quad \#S \leq s\,, \quad C\|h_S\|_1 \leq \|h_{S^c}\|_1\,,$$

where $\#S$ denotes the size of the set S, S^c denotes the complement of S, and the vector h_S equals h on the set of entries S and is null otherwise.

As a matter of fact, one can prove [CDD09] that the operator $\mathbf{D}_X$ is the identity on Σ_s if and only if the kernel of X satisfies NSP(s, C) for some $C > 1$.

Theorem 1 ([CDD09])**.** *For all $\beta^0 \in \Sigma_s$ there is a unique solution to (1.1) and $\mathbf{D}_X(\beta^0) = \beta^0$ if and only if the null space* $\ker(X)$ *of the matrix X enjoys* NSP(s, C) *for some $C > 1$. In this case, for all $\beta^0 \in \mathbb{R}^p$,*

$$\|\beta^0 - \mathbf{D}_X(\beta^0)\|_1 \leq \frac{2(C+1)}{C-1}\sigma_s(\beta^0)_1\,.$$

where $\sigma_s(\beta^0)_1$ denotes the ℓ_1-approximation error by Σ_s, namely $\sigma_s(\beta^0)_1 := \min\|\beta^0 - \beta\|_1$ where the minimum is taken over the space Σ_s of s-sparse vectors β.

It shows that one can exactly recover any sparse vector by ℓ_1-minimization, which is referred to as the "*Exact Reconstruction property*" [CGLP12, Definition 2.2.10]. Additionally, NSP suffices to show that any solution to (1.1) is comparable to the s-best approximation $\sigma_s(\beta^0)_1$ of the target vector β^0. Theorem 1 demonstrates that NSP is a natural property that should be required in CS and high-dimensional Statistics.

This analysis can be taken a step further considering the Lasso estimator [Tib96] or the Dantzig selector [CT07], see Section 1.2. We will see that, in the framework of noisy observations, the ℓ_1-minimization procedures are based on sufficient conditions such as the Restricted Isometry Property (RIP) [CT06, CT07] for instance. Note that all of these properties imply that the kernel of the matrix X satisfies NSP. While there exists pleasingly ingenious and simple proofs of RIP, see [CGLP12] for instance, a direct proof of NSP (without the use of RIP) remains a challenging issue.

1.1.2 Proving the Null Space Property

Few works achieve a direct proof of NSP. They are based either on integral convex geometry theory [DT05, Don06b, DT09a, DT09b], Gaussian widths [Sto10, Sto13], the approximate kinematic formula [MT14, ALMT14], empirical process theory [LM16], or suprema of piecewise linear Gaussian processes **[DC9]**. Interestingly, Donoho and Tanner [DT05, Don06b, DT09a] have proved that random projections of the s-faces of the cross polytope satisfy a "*phase transition*". In particular, it yields [Don05] that there exists a function $\rho_S\colon]0,1[\to]0,1[$ such that for all $(\rho, \delta) \in]0,1[^2$,

$$\lim_{n\to+\infty} \mathbb{P}[\ker(X(n, p_n)) \text{ enjoys NSP}(s_n, 1)] = \begin{cases} 0 & \text{if } \rho > \rho_S(\delta) \\ 1 & \text{if } \rho < \rho_S(\delta) \end{cases},$$

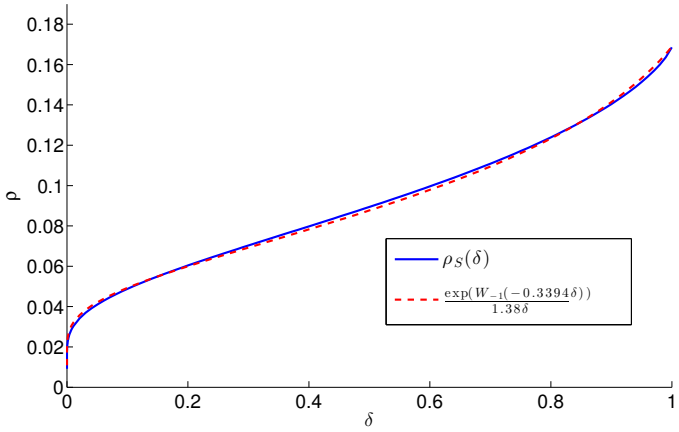

Figure 1.1: The strong threshold ρ_S and an approximation by Lambert functions.

where $s_n = \lfloor \rho\, n \rfloor$, $p_n = \lfloor n/\delta \rfloor$ and the design matrix $X(n, p_n) \in \mathbb{R}^{n \times p_n}$ has i.i.d. centered Gaussian entries. Moreover, they have characterized implicitly and computed numerically the function[1] ρ_S, see Figure 1.1 for an illustration.

On the other hand, the bound (1.2) is a striking result of Compressed Sensing and one can wonder up to what extend it compares to the strong threshold ρ_S that describes NSP with dilatation $C = 1$. As mentioned in [CGLP12, Proposition 2.2.18], if NSP holds then (1.2) holds. The result of [LM16] shows[2] that the same bound (up to change of constants) is also sufficient to get NSP. What can be understood is that the phase transition lies between two bounds described by (1.2). Observe that these bounds can be equivalently expressed in terms of $\rho := s/n$ and $\delta := n/p$. Indeed, one has

$$\left\{ n \geq c_1 s \log\left(\frac{c_2 p}{s}\right) \right\} \Leftrightarrow \left\{ A_\star \rho\, \delta \log(A_\star \rho\, \delta) \geq -B_\star \delta \right\}, \tag{1.3}$$

where $A_\star = c_2^{-1} > 0$ and $1/e \geq B_\star = c_1^{-1} c_2^{-1} > 0$. Denote by $\mathbf{W}_{-1}$ the second Lambert W function, see [CGH+96] for a definition[3]. From (1.3), one can deduce the following result.

Proposition 2. *The strong threshold ρ_S of Donoho and Tanner is bounded by*

$$\forall \delta \in (0,1), \quad \frac{\exp(\mathbf{W}_{-1}(-B_1\delta))}{A_1\delta} \leq \rho_S(\delta) \leq \frac{\exp(\mathbf{W}_{-1}(-B_2\delta))}{A_2\delta}$$

where $A_1, A_2 > 0$ and $1/e \geq B_1, B_2 > 0$ are universal (unknown) constants.

As a matter of fact, Figure 1.1 depicts a comparison between ρ_S and

$$\delta \mapsto \frac{\exp(\mathbf{W}_{-1}(-0.3394\delta))}{1.38\delta}, \tag{1.4}$$

where the strong threshold curve has been taken from [DT05, Don06b, DT09a]. Roughly speaking, the curve (1.4) shows empirically that NSP holds when $n \geq 4s \log(0.7p/s)$ for large values of s, n, p. Recall that it is still an open problem to find a closed form for the weak and the strong thresholds. In the regime $\delta \to 0$, Donoho and Tanner [DT05, Don06b, DT09b, DT09a] have proved that the phase transition enjoys $n \geq 2e\, s \log(p/(\sqrt{\pi} s)) \simeq 5.4 s \log(0.6p/s)$ in the asymptotic.

Using integral convex geometry theory as in Donoho and Tanner's works, Xu and Hassibi have investigated [XH08, XH11] the property NSP(s, C) for values $C > 1$. Their result uses an implicit equation involving inverse Mill's ratio of the normal distribution. To the best of our knowledge, this is the only proof of NSP(s, C) for values $C > 1$ predating **[DC9]**. Indeed, using a Kac-Rice formula on piecewise regular Gaussian processes, one can provide a new description of the region where NSP holds for parameters (ρ, δ) as follows.

[1] Note that the subscript S stands for "Strong" since ρ_S is often named the "strong threshold".
[2] We point to this reference since it furnishes a direct proof of NSP.
[3] Lambert functions are the inverses of $x \mapsto x \exp x$ and $\mathbf{W}_{-1}$ the inverse defined on the branch $x < 0$.

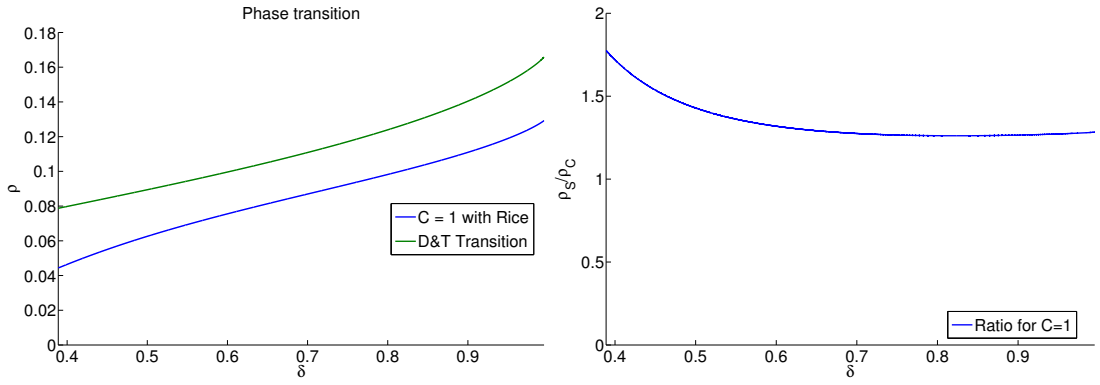

Figure 1.2: Comparison of ρ_C given by Theorem 3 with the strong threshold ρ_S.

Theorem 3 ([DC9]). *Let $C \geq 1$. If $\delta \geq (1+\pi/2)^{-1} \simeq 0.389$ and*

$$\rho \log\left(\sqrt{\frac{\pi}{2eC^2}}\frac{(1-\rho)^2}{\rho^2}\right) + \log\left(Ce\frac{\sqrt{\rho(1-\delta)(1+(C^2-1)\rho)}}{(1-\rho)(1+(2C^2-1)\rho)\sqrt{\delta}}\right) + \frac{1}{\delta}\log\left(\sqrt{\frac{2}{e\pi}}\frac{1+(2C^2-1)\rho}{(1-\rho)\sqrt{\delta(1-\delta)(1+(C^2-1)\rho)}}\right) \leq 0$$

then $\mathbb{P}[\ker(X(n,p_n))$ enjoys $\mathrm{NSP}(s_n,1)]$ tends exponentially fast to one as n goes to infinity. Here, $s_n = \lfloor \rho\, n \rfloor$, $p_n = \lfloor n/\delta \rfloor$ and the design matrix $X(n,p_n) \in \mathbb{R}^{n\times p_n}$ has i.i.d. centered Gaussian entries.

Key step(s) of the proof: The proof is based on the Rice method [AW09, Chapter 3] for a non differentiable and non Gaussian process defined on the sphere. Our argument uses a partition of this sphere and applies the Rice method on pieces of the partition. Summing up we obtain lower bound on the event "NSP holds". The number of pieces of our partition decreases with $\delta = n/p$ so our method is better for δ bounded away from zero. ■

Remark 1. One can compare the result given by Theorem 3 to the work of Donoho and Tanner [DT05]. Indeed, in the case $C = 1$, the "lower bound" on NSP given by Theorem 3 is the region $(\rho,\delta) \in]0,1[^2$ such that $\delta \geq (1+\pi/2)^{-1}$ and

$$\rho \log\left[\sqrt{\frac{\pi}{2e}}\frac{(1-\rho)^2}{\rho^2}\right] + \log\left[e\frac{\sqrt{\rho(1-\delta)}}{(1-\rho)(1+\rho)\sqrt{\delta}}\right] + \frac{1}{\delta}\log\left[\sqrt{\frac{2}{e\pi}}\frac{1+\rho}{(1-\rho)\sqrt{\delta(1-\delta)}}\right] \leq 0,$$

as depicted in Figure 1.2. Observe that, up to a multiplicative constant bounded by 1.8, we recover the strong phase transition on NSP.

1.2 Stable and Robust Sparse Recovery

1.2.1 Universal Distortion Property

From now on, consider the linear model with stochastic error term ζ given by $y = X\beta^0 + \zeta$ where we recall that X is a known $n \times p$ matrix, β^0 a unknown vector in $\mathbb{R}^p$, y and ζ are vectors in $\mathbb{R}^n$ and n is (much) smaller than p. Although the matrix X is not injective, we have seen in Section 1.1.2 that one can recover an interesting estimate $\widehat{\beta}$ of β^0 using ℓ_1-minimization solutions. In the case of noisy observations, one can consider an estimator $\widehat{\beta}$ given by

$$\widehat{\beta} \in \arg\min \|\beta\|_1 \quad \text{s.t.} \quad \|y - X\beta\|_n \leq \eta \tag{1.5}$$

where $\eta > 0$ is a tuning parameter and $\|\cdot\|_n = n^{-\frac{1}{2}}\|\cdot\|_2$. Then, an appealing goal is to prove that, with high probability, it holds that

$$\|\beta^0 - \widehat{\beta}\|_1 \leq C\sigma_s(\beta^0)_1 + D\sqrt{s}\eta \tag{1.6}$$

$$\|\beta^0 - \widehat{\beta}\|_2 \leq \frac{C}{\sqrt{s}}\sigma_s(\beta^0)_1 + D\eta \tag{1.7}$$

where $C, D > 0$ are constants. The important feature described by (1.6) and (1.7) may be referenced to as the "*Stable and Robust Sparse Recovery*" (SRSR) property of order s, see [FR13, Page 88]. Roughly speaking, it shows that ℓ_1-minimization recovers the s largest coefficients of a target vector β^0 in a stable[4] and robust (to additive errors ζ) manner.

Using the SRSR view point, standard results [FR13, vdG16] can be stated for estimators such as the Lasso [Tib96] or the Dantzig selector [CT07]. More precisely, in view of (1.7), one may require that η in (1.6) and (1.7) is the minimax estimation rate $r_{n,p}(\Sigma_s)$ for the ℓ_2-loss [RWY11], that is to say

$$\eta = r_{n,p}(\Sigma_s) := c_0\sigma\sqrt{s\log\big(p/s\big)/n}\,,$$

where σ is the standard deviation parameter on the Gaussian noise ζ and $c_0 > 0$ some numerical constant. To get such results, recall that the Lasso is given by

$$\widehat{\beta} \in \arg\min_{\beta\in\mathbb{R}^p}\Big\{\frac{1}{2}\|y - X\beta\|_n^2 + \lambda\|\beta\|_1\Big\} \tag{1.8}$$

where $\lambda > \lambda_0 := n^{-1}\|X^\top\zeta\|_\infty$ with high probability[5]. In high-dimensional Statistics, one usually assumes that $\max_{k\in[p]}\|X_k\|_n$ is upper bounded by a constant where $(X_k)_{k\in[p]}$ denotes the columns of X. This assumption together with the forthcoming UDP property (1.9) lead to a so-called "*oracle inequality*" given by (1.6) with $\eta := r_{n,p}(\Sigma_s)$, see Theorem 4. More generally, for the Lasso, the Dantzig selector or (1.5), it has been proved that SRSR holds whenever the matrix X satisfies some properties, see for instance [CRT06, CT06, FL09, BRT09, vdGB09, BLPR11, JN11] or [CGLP12, FR13, Gir14] for valuable books.

Let us comment on these different conditions on X proving SRSR. In Section 1.2.2, we will consider the different offsprings of the Restricted Isometry Property (RIP) that have attracted a lot of attention in the past decade. As for now, let us focus on the weakest condition to get SRSR. Following **[DC2]**, the Universal Distortion Property of order $s \in [n]$, magnitude $\kappa_0 \in (0, 1/2)$ and parameter $\Delta > 0$, referred to as UDP(s, κ_0, Δ), is defined by[6]

$$\forall h \in \mathbb{R}^p\,, \quad \forall S \subset [p] \quad \text{s.t.} \quad \#S \leq s\,, \quad \|h_S\|_1 \leq \kappa_0\|h\|_1 + \Delta\sqrt{s}\|Xh\|_n\,. \tag{1.9}$$

One can prove **[DC2]** that UDP is a weaker requirement on the design X than RIP [CT06], Restricted Eigenvalue [BRT09], Compatibility [vdGB09, BvdG11], Irrepresentability [Fuc05], $\mathbf{H}_{s,1}$ [JN11], or ℓ_2-robust null space property[7] [FR13]. Furthermore, we have the following result.

Theorem 4 (**[DC2]**)**.** *Assume that the design X satisfies* UDP(s, κ_0, Δ) *and assume also that* $\max_{k\in[p]}\|X_k\|_n \leq 1$. *Then the SRSR property* (1.6) *holds where $\widehat{\beta}$ is the solution to Lasso, Dantzig selector or* (1.5). *Namely, for the Lasso and the Dantzig selector, it holds that*

$$\|\beta^0 - \widehat{\beta}\|_1 \leq C\sigma_s(\beta^0)_1 + Ds\sigma\sqrt{\log p/n}\,,$$

with $C, D > 0$ numerical constants and a choice $\lambda = c_1\sigma\sqrt{\log p/n}$ as tuning parameter (1.8) *for some numerical $c_1 > 0$.*

[4] In a idealized situation one would assume that β^0 is sparse. Nevertheless, in practice, we can only claim that β^0 is close to sparse vectors. The stability is the ability to control the estimation error $\|\beta^0 - \widehat{\beta}\|$ by the distance between β^0 and the sparse vectors. The reader may consult [FR13, Page 82] for instance.

[5] One may choose $\lambda_0 \geq c\,\max_{k\in[p]}\|X_k\|_n\sigma\sqrt{\log p/n}$ with some numerical constant $c > 0$.

[6] In **[DC2]**, note that the UDP property is stated with $\|X\beta\|_2$ and here with $\|X\beta\|_n$. We choose this latter formulation so that the UDP constant Δ is "*homogeneous*" (*i.e.,* same dependency in s and n) with the Compatibility constant [vdG16, Section 2.6] and the Restricted Eigenvalue constant [BRT09, Section 3].

[7] Setting $\rho = \kappa_0/(1-\kappa_0)$ and $\tau = \Delta/(1-\kappa_0)$ and noticing that $s^{-\frac{1}{2}}\|\beta_S\|_1 \leq \|\beta_S\|_2$, one can readily prove that UDP is implied by the ℓ_2-robust null space property as defined in [FR13, Definition 4.21].

A similar result holds for minimax prediction error for the Lasso and the Dantzig selector using the UDP property, see **[DC2]**. The dependence of the constants C and D in κ_0 and Δ are given by $C = 2/[(1-\lambda_0/\lambda)-2\kappa_0]$ and $D = 2c_1\Delta^2/[(1-\lambda_0/\lambda)-2\kappa_0]$.

Key step(s) of the proof: For the Lasso, the argument is standard and results in comparing the objective function at points β^0 and $\widehat{\beta}$. It yields

$$\frac{1}{2\lambda}\Big[\frac{1}{2}\|Xh\|_n^2 + (\lambda-\lambda_0)\|h\|_1\Big] \le \|h_S\|_1 + \sigma_s(\beta^0)_1$$

where $\lambda > \lambda_0 := n^{-1}\|X^\top\zeta\|_\infty$ and $h = \widehat{\beta}-\beta^0$. Invoke UDP to get

$$\Big[\frac{1}{2}\Big(1-\frac{\lambda_0}{\lambda}\Big)-\kappa_0\Big]\|h\|_1 \le \underbrace{-\frac{1}{4\lambda}\Big[\|Xh\|_n^2 - 4\lambda\Delta\sqrt{s}\|Xh\|_n\Big]}_{\le \Delta^2 s\,\lambda} + \sigma_s(\beta^0)_1\,.$$

using the fact that $\{x^2 - bx \le c\}$ implies $\{x \le b + c/b\}$. The interested reader may find an other proof in [FR13, Theorem 4.25], noticing that [FR13, Eq. (4.18)] is the UDP property. ■

The UDP property is intimately related to the "*distortion*[8]" δ of the kernel of the design X. Precisely, we recall that it has been established [Kaš77] that, with an overwhelming probability, a random[9] subspace $\Gamma_n \subset \mathbb{R}^p$ of dimension $p-n$ satisfies

$$\delta \le \delta^\star_{n,p} := C\left(\frac{p(1+\log(p/n))}{n}\right)^{1/2}, \tag{1.10}$$

where $C > 0$ is some constant. In other words, it was shown that, for all $\beta \in \Gamma_n$, it holds that $\|\beta\|_1 \le \sqrt{p}\|\beta\|_2 \le \delta^\star_{n,p}\|\beta\|_1$.

Theorem 5 ([DC2])**.** *Let $X \in \mathbb{R}^{n\times p}$ be a full rank matrix. Denote by δ the distortion of its kernel and ρ_n the smallest singular value of $X/\sqrt{n}$. Let $\kappa_0 \in (0,1/2)$ then X satisfies* UDP(s,κ_0,Δ) *with parameters $s := (\kappa_0/\delta)^2 p$ and $\Delta := 2\delta/\rho_n$.*

This theorem is sharp in the following sense. The parameter s represents the maximum number of coefficients that can be recovered as shown by Theorem 4. From (1.2) in Section 1.1.2, one knowns that the best bound one could expect is $s^\star \simeq n/\log(p/n)$, up to a log factor. In the case where (1.10) holds and in view of $s := (\kappa_0/\delta)^2 p$, the sparsity level satisfies $s \simeq \kappa_0^2 s^\star$. It shows that any design matrix with low distortion satisfies UDP with an optimal sparsity level. Remark that small distortion kernels can be achieved using designs with i.i.d. rotationally invariant lines.

Furthermore, a standard result [BS10, Tik15] shows that ρ_n almost surely converges towards the square root of the left border of the Marchenko-Pastur law when X has i.i.d. centered unit variance entries. An example of UDP matrix is given by random matrices $X \in \mathbb{R}^{n\times p}$ with i.i.d.Rademacher entries (or any unit variance sub-Gaussian law) with $n \ge c\,s\log(c\,p/n)$ and $c > 0$ a numerical constant. Then, one can show **[DC7]** that X satisfies UDP$(s,4/9,9/2)$ with high probability.

Last but not least, UDP is also relevant for some deterministic designs. In particular one can prove **[DC3]** that the normalized adjacency matrix[10] of an (s,ε)-unbalanced expander graph with expansion constant ε less than $1/12$ satisfies UDP$(s,1/5,6/5)$ and, by Theorem 4, the SRSR property (1.6). One can also prove **[DC3]** that minimax prediction error holds using those matrices with the Lasso or the Dantzig selector. Note that expander design adjacency matrices can be deterministically constructed based on Paravaresh-Vardy codes [PV05] in polynomial time [GUV09].

[8] Maximal ratio between the $\sqrt{p}\,\|\cdot\|_2$ and the $\|\cdot\|_1$.

[9] with respect to the Haar measure on the Grassmannian.

[10] Following the exposition of this dissertation, one has to consider the normalization $X := C\sqrt{s}\,\Phi$ where $C > 0$ is a constant, Φ is the adjacency matrix, and s is an expansion parameter that is known here.

1.2.2 Restricted Isometry Constants

One of the most important properties in high-dimensional Statistics is undoubtedly the Restricted Isometry Property[11] [CRT06, CT06] of order $s \in [n]$ and RIP constant $c \in (0,1)$, referred to as RIP(s,c). One can prove [FR13, Theorem 6.12] that, if RIP(s,c) with RIP constant $c < 4/\sqrt{41} \simeq 0.625$ and $\widehat{\beta}$ is any solution to (1.5) then the SRSR property of order s holds with constants C, D depending only on c. A similar result holds for the Lasso and the Dantzig selector. A slightly modified RIP was introduced by Foucart and Lai in [FL09] under the notion of Restricted Isometry Constants (RICs).

Definition (Restricted Isometry Constants (RICs))**.** *For a matrix X of size $n \times p$, the restricted isometry constants (RICs) $\mathfrak{c}_{\min}(s,X)$ and $\mathfrak{c}_{\max}(s,X)$ are defined as*

$$\mathfrak{c}_{\min} := \min_{c_- \geq 0} c_- \quad \text{subject to} \quad (1-c_-)\|\beta\|_2^2 \leq \|X\beta\|_2^2 \quad \text{for all} \quad \beta \in \Sigma_s,$$
$$\mathfrak{c}_{\max} := \min_{c_+ \geq 0} c_+ \quad \text{subject to} \quad (1+c_+)\|\beta\|_2^2 \geq \|X\beta\|_2^2 \quad \text{for all} \quad \beta \in \Sigma_s.$$

Hence, it holds that $(1-\mathfrak{c}_{\min})\|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq (1+\mathfrak{c}_{\max})\|\beta\|_2^2$ for all $\beta \in \Sigma_s$, where we recall that Σ_s denotes the set of vectors with at most s nonzero coordinates.

Interestingly, Foucart and Lai proved the following result.

Theorem 6 (Theorem 2.1 in [FL09])**.** *If X satisfies*

$$\frac{1+\sqrt{2}}{4}\left[\frac{1+\mathfrak{c}_{\max}}{1-\mathfrak{c}_{\min}} - 1\right] < 1, \tag{1.11}$$

then the Stable and Robust Sparse Recovery property (1.6) and (1.7) of order s holds with positive constants C and D depending only on $\mathfrak{c}_{\min}(2s,X)$ and $\mathfrak{c}_{\max}(2s,X)$.

One engaging feature of Condition (1.11) is that it reports the influence of both extreme eigenvalues of covariance matrices built from $2s$ columns of X.

In **[DC16]**, we provide a simple tool to derive a region of parameters s, n, p for which SRSR holds, and also upper bounds on RICs. Our point of view is to use deviation inequalities on extreme eigenvalues (or singular values) of covariance matrices $\mathbf{C}_{s,n} = \frac{1}{n}\mathbf{X}\mathbf{X}^\star$ where the matrix $\mathbf{X} \in \mathbb{R}^{s \times n}$ has i.i.d. entries drawn with respect to $\mathscr{L}$. In the asymptotic proportional growth model where $s/n \to \rho$ and $n/p \to \delta$, we assume that we have access to a deviation inequality on extreme eigenvalues with rate function $t \mapsto \mathbb{W}(\rho, t)$ depending on the ratio ρ. For instance, we will consider that for all $n \geq n_0(\rho)$,

$$\forall 0 \leq t < \tau_1, \quad \mathbb{P}\left\{\left(\lambda_1 - (1+\sqrt{\rho})^2\right) \vee \left((1-\sqrt{\rho})^2 - \lambda_s\right) \geq t\right\} \leq c(\rho) e^{-n\mathbb{W}(\rho,t)}$$

where $\tau_1 \in \overline{\mathbb{R}}$, $n_0(\rho) \geq 2$ and $c(\rho) > 0$ may both depend on the ratio ρ, the function $t \mapsto \mathbb{W}(\rho,t)$ is continuous and increasing on $[0, \tau_1)$ such that $\mathbb{W}(\rho,0) = 0$, and λ_1 (resp. λ_s) denotes the smallest (resp. largest) eigenvalue of the Wishart matrix $\mathbf{C}_{s,n}$. Notably, it appears throughout our analysis that the upper bounds on RICs are extremely dependent on the behavior, for fixed t, of the function $\rho \mapsto \mathbb{W}(\rho,t)$ when ρ is small, and possibly tending to zero. Unfortunately, this dependence is overlooked in the literature and we have to take another look at state-of-the-art results in this field. Revisiting the captivating paper of Feldheim and Sodin [FS10] on sub-Gaussian matrices, **[DC16]** reveals the dependency on ρ as well as bounds on the constant appearing in their rate function $\mathbb{W}_{FS}$ for the special case of Rademacher entries. Other important rate functions due to Ledoux and Rider [LR+10], and Davidson and Szarek [DS01] are investigated in **[DC16]**.

[11] Recall its expression: $\forall \beta \in \Sigma_s$, $(1-c)\|\beta\|_2^2 \leq \|X\beta\|_2^2 \leq (1+c)\|\beta\|_2^2$.

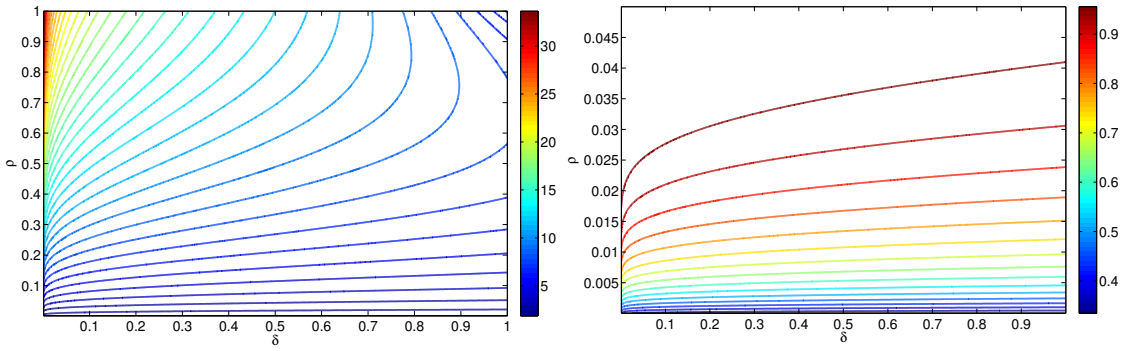

Figure 1.3: Davidson and Szarek bounds on RICs.

Theorem 7 ([DC16]). *The following holds for each couple* $(\mathbb{W},\mathscr{L})$ *defined by, for all* $t>0$, *for all* $\rho\in(0,1)$,

$$\mathbb{W}_{LR}(\rho,t):=\frac{\rho^{\frac{1}{4}}}{C_{LR}(1+\sqrt{\rho})^3}t^{\frac{3}{2}}\mathbb{1}_{t\leqslant\sqrt{\rho}(1+\sqrt{\rho})^2}+\frac{\rho^{\frac{1}{2}}}{C_{LR}(1+\sqrt{\rho})^2}t\mathbb{1}_{t>\sqrt{\rho}(1+\sqrt{\rho})^2} \quad \text{and} \quad \mathscr{L}_{LR}:=\mathcal{N}(0,1)$$

$$\mathbb{W}_{FS}(\rho,t):=\frac{\rho\log(1+\frac{t}{2\sqrt{\rho}})^{\frac{3}{2}}}{C_{FS}(1+\sqrt{\rho})^2} \quad \text{and} \quad \mathscr{L}_{FS}:=\text{Rademacher}$$

where $C_{LR}>0$ *and* $837\geq C_{FS}>0$ *are numerical constants.*
∘ *For any* $\varepsilon>0$, *any* $\delta\in(0,1)$ *and any* $\rho\in(0,1)$, *it holds*

$$\mathbb{P}\Big\{\mathfrak{c}_{\min}\geqslant\min\{1,\sqrt{\rho}(2-\sqrt{\rho})+t_0\}+\varepsilon\Big\}\leqslant c(\rho)e^{-nD(\rho,\delta,\varepsilon)},$$
$$\mathbb{P}\Big\{\mathfrak{c}_{\max}\geqslant\sqrt{\rho}(2+\sqrt{\rho})+t_0+\varepsilon\Big\}\leqslant c(\rho)e^{-nD(\rho,\delta,\varepsilon)},$$

where $D(\rho,\delta,\varepsilon)>0$ *and* $t_0:=\mathbb{W}^{-1}(\rho,\delta^{-1}\mathbf{H}_e(\rho\delta))$.
∘ *Let* $\rho_0:=(3+\sqrt{2}-\sqrt{7+6\sqrt{2}})^2/4\simeq 0.0574$ *and* $\tau_0:=2/(3+\sqrt{2})\simeq 0.4531$. *For any* $\delta\in(0,1)$ *and any* $\rho\in(0,\rho_0)$ *such that*

$$\delta>\frac{1}{\rho}\exp\left[1-\frac{\mathbb{W}\big[\rho,\tau_0(\sqrt{\rho}-\sqrt{\rho_0})(\sqrt{\rho}-1/\sqrt{\rho_0})\big]}{\rho}\right], \tag{1.12}$$

it holds that any sequence of $n\times p$ *matrices* $(X^{(n)})_{n\geq 2}$ *with i.i.d. entries with respect to* $\mathscr{L}$ *and such that* $n/p\to\delta$ *satisfy* $\mathbb{P}\Big\{\frac{X^{(n)}}{\sqrt{n}}$ conforms to (1.11) with $2s\leq\lfloor\rho n\rfloor\Big\}\to 1$ *as* n *tends to infinity.*

Key step(s) of the proof: First, we aim at controlling uniformly the extreme eigenvalues, the combinatorial complexity is standardly (see [CT05, Lemma 3.1] or [BCT11]) given by the quantity $\delta^{-1}\mathbf{H}_e(\rho\delta)$ where $\mathbf{H}_e(t)=-t\log t-(1-t)\log(1-t)$ for $t\in(0,1)$ denotes the Shannon entropy. Then, this quantity governs the value of the deviation $t_0:=\mathbb{W}^{-1}(\rho,\delta^{-1}\mathbf{H}_e(\rho\delta))$ in the rate function $\mathbb{W}(\rho,t)$ when bounding the extreme eigenvalues uniformly over all possible supports S of size s among the set of indices $[p]$. ■

Observe that (1.12) describes a region (ρ,δ) (referred to as the "*lower bound*") for which the SRSR properties (1.6) and (1.7) hold. Using the rate function of Davidson and Szarek [DS01], [DC16] derives an upper bounds on RICs (Fig. 1.3) and a lower bound on SRSR (Fig. 1.4).

Another work looking at "*phase transition*" on SRSR can be found in the captivating paper [BCT11] where the authors considered matrices with independent standard Gaussian entries and used an upper bound on the joint density of the eigenvalues to derive a region

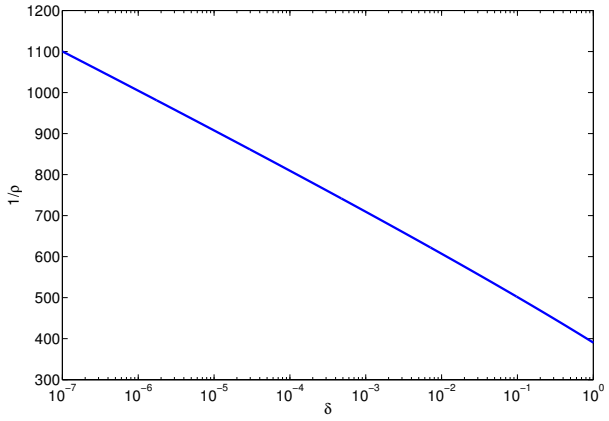

Figure 1.4: Lower bound on SRSR using Davidson and Szarek deviations.

where (1.11) holds. Their lower bound is not explicit but one can witness it in their paper [BCT11, Page 119]. The lower bound presented in Figure 1.4 is comparable to this latter bound up to a multiplicative constant smaller than 2.

1.2.3 Sensitivity Analysis

In **[DC7]**, we investigate selection of variables in high-dimension from a nonparametric regression model depicting a situation where we are concerned with estimating a nonparametric regression function f that may depend on a large number p of input variables. Unlike standard procedures, we do not assume that f belongs to a class of regular functions, yet we assume that f is a square-integrable function with respect to a known product measure. We assume that only a small number s of the coordinates actually affects f in an additive manner. In this frame, we prove that, with only $\mathcal{O}(s\log p)$ random evaluations of f, one can find which are the relevant input variables (referred to as the "*support*") with high probability. Our proposed method is an unconstrained ℓ_1-minimization procedure based on Sobol's method. One step of this procedure relies on the "*thresholded*"-Lasso to faithfully uncover the significant input variables. We won't present Sobol's method here nor how to rephrase it in a high-dimensional regression fashion—the interested reader may consult **[DC7]**. However, we recall here the key results of high-dimensional Statistics that are involved in the so-called "Randomized Pick-Freeze" (RPF) method of **[DC7]**.

From the high-dimensional Statistics perspective, the RPF method consists in support recovery using the thresholded-Lasso when the design matrix $X \in \mathbb{R}^{n\times p}$ is drawn with respect to Bernoulli or Rademacher laws. The thresholded-Lasso is a standard procedure to estimate the support of a sparse vector from few linear measurements. It is based on two features: an assumption that the nonzero coefficients of the target $\beta^0 \in \Sigma_s$ have a magnitude greater than a threshold[12] $\tau_0 = c\lambda$ for some numerical constant $c > 0$, and a control of the "*mutual incoherence*" property [DET06] of the form

$$\max_{1\le k\neq \ell\le p} \frac{1}{n}|\sum_{j=1}^{n} X_{j,k}X_{j,\ell}| \le \frac{1}{2s-1}\min_{k\in[p]}\|X_k\|_n^2 .$$

Using Welch's bound [Wel74], one can prove that this condition implies $n \ge C\,s^2\log p$ for some constant $C > 0$. An other condition that can be used is the "*Irrepresentability Condition*" (IC) [Fuc04, ZY06] though this approach is rather stringent and proving (IC) for random matrices remains a challenging issue. We now present a new approach, based on UDP (see Section 1.2.1) and a relaxed version of the coherence which enables to break the s^2-"*bottleneck*" for a large set of random design matrices encompassing Rademacher designs, see **[DC7]**.

[12] where λ denotes the tuning parameter of the Lasso (1.8).

Theorem 8 ([DC7]). *Assume that X satisfies* $\mathrm{UDP}(s,\kappa_0,\Delta)$ *for some* $\kappa_0 \in (0,1/2)$. *On the event defined by*

$$\max_{1\le k\neq \ell\le p} \frac{1}{n}\Big|\sum_{j=1}^{n} X_{j,k}X_{j,\ell}\Big| \le \theta_1 \quad \text{and} \quad \forall k\in[p], \quad \theta_2 \le \|X_k\|_n^2 \le 1,$$

and $\lambda_0 \ge n^{-1}\|X^\top \zeta\|_\infty$, *the following holds. If* λ *in* (1.8) *such that* $\lambda \ge (1-2\kappa_0)^{-1}\lambda_0$, *then the Lasso estimator* (1.8) *satisfies*

$$\|\widehat{\beta}-\beta^0\|_\infty \le \frac{1}{\theta_2}\Big[1+\frac{\lambda_0}{\lambda}+\frac{2\theta_1\Delta^2 s}{1-(\lambda_0/\lambda)-2\kappa_0}\Big]\lambda.$$

Key step(s) of the proof: The proof uses "*ad hoc*" Gram matrix concentration and follows the same guidelines as standard proof [Lou08] on the thresholded-Lasso while invoking the ℓ_1-error control given by UDP, see Theorem 4. ■

This theorem can be invoked for matrices with i.i.d. sub-Gaussian entries, *e.g.,* Rademacher entries. It gives the following standard result, one may also consult [CP09, Wai09].

Corollary 1. *Let* $X \in \mathbb{R}^{n\times p}$ *be a random matrix with i.i.d. unit variance sub-Gaussian entries. There exist numerical constants* $c_0, c_1 > 0$ *such that the following holds. If* $n \ge c_0 s\log p$ *and* $\lambda \ge c_0\sigma\sqrt{\log p/n}$ *then the solution* $\widehat{\beta}$ *to the Lasso* (1.8) *satisfies*

$$\|\widehat{\beta}-\beta^0\|_\infty \le c_1\sigma s\sqrt{\log p/n},$$

with high probability.

Key step(s) of the proof: In this case, one can choose, for any $\alpha \in [0,1]$, θ_1 of the order of $s^{-\frac{1+\alpha}{2}}$, n of the order of $s^{1+\alpha}\log p$ and $\theta_2 > 0$ constant. It gives $\|\widehat{\beta}-\beta^0\|_\infty \le c_1 s^{-\frac{1-\alpha}{2}}\lambda$. ■

In particular, choosing a threshold $\tau_0 := c_1\lambda\sqrt{s^2\log p/n}$ and assuming that the nonzero coefficients of β^0 are greater (in absolute value) than $2\tau_0$, we deduce that the thresholded-Lasso (Lasso with a hard-thresholding step) faithfully recovers the target support with high probability. Notice that in the case $n \ge c_0 s^2\log p$, we recover the standard results in the literature that is to say the threshold is given by $\tau_0 := c\lambda$, for some numerical constant $c > 0$.

1.3 Exact Post-Selection Inference

Recent advances have focused on hypothesis testing using penalized problems, see for instance [LSST13, LTTT14, TLT13, TLTT14] and references therein. Compared to the sparse recovery problems, very little work has been done in statistical testing in high dimensions. As a matter of fact, one of the main difficulties is that there is no tractable distribution of sparse estimators (even under the RIP-like standard conditions of high-dimensional Statistics). A successful approach is then to take into account the influence of each predictor in the regression problem. More precisely, some recent works in "*Post-Selection Inference*" have shown that the selection events can be explicitly[13] expressed as closed convex polytopes depending simply on the signs and the indices of the nonzero coefficients of the solutions of standard procedures in high-dimensional Statistics (typically the solutions of the Lasso). Furthermore, an important advance has been brought by a useful parametrization of these convex polytopes under the Gaussian linear model, see for instance [HTW15, Chapter 6.3.2]. In detection testing, this is done by the first two "*knots*" of the *least-angle regression algorithm* (LARS for short) which is intimately related to the dual program of the ℓ_1-minimization problem, see [EHJT04] for example.

[13] using KKT conditions of ℓ_1-minimization programs, for instance.

In the Gaussian linear model, further works have derived unconditional test statistics such as the so-called "*Kac-Rice Pivot*" for general penalized problems. In order to test the global null, a prominent offspring of this breakthrough is the "*Spacing test*" that accounts the relative separation between consecutive knots of the LARS. However, no results have been obtained regarding the distribution of these test statistics under the alternative. In **[DC10]**, we address this important issue for the spacing test and we show that it is unconditionally unbiased. Furthermore, we provide an extension of the spacing test to the frame of unknown noise variance.

1.3.1 Hypothesis testing using LARS

In this section, we consider an outcome vector $y \in \mathbb{R}^n$, a matrix of predictor variables $X \in \mathbb{R}^{n\times p}$ and a variance-covariance matrix Θ such that

$$y = X\beta^0 + \zeta \quad \text{with} \quad \zeta \sim \mathcal{N}_n(0,\Theta).$$

We are concerned with testing whether β^0 is equal to some known β_0^0 or not. Notice that the response variable y does not depend directly on β^0 but rather on $X\beta^0$ which is known. Subtracting $X\beta_0^0$, a detection test may be interested in discerning between two hypothesis on the target vector β^0, namely $\mathbb{H}_0$: "$\beta^0 \in \ker(X)$" against $\mathbb{H}_1$: "$\beta^0 \notin \ker(X)$". To this end, we consider the vector of correlations $U := X^\top y \sim \mathcal{N}_p(\mu^0, R)$ where $\mu^0 := X^\top X\beta^0$ and $R := X^\top \Theta X$. Observe that the hypotheses $\mathbb{H}_0$ and $\mathbb{H}_1$ can be equivalently written as

$$\mathbb{H}_0 : "\mu^0 = 0" \quad \text{against} \quad \mathbb{H}_1 : "\mu^0 \neq 0",$$

and remark that the knowledge of the noise variance-covariance matrix Θ is equivalent to the knowledge of the correlations variance-covariance matrix R.

1.3.2 Power of the spacing test for LARS

The test statistic we are considering was introduced in a larger context of penalization problems by the pioneering works in [TLTT14, TLT13]. As mentioned by the authors of [TLT13], the general test statistic "*may seem complicated*". However, it can be greatly simplified in the frame of the standard regression problems under a very mild assumption, namely

$$\forall i \in [p], \quad R_{ii} := X_i^\top \Theta X_i = 1. \tag{1.13}$$

Note that this assumption is not very restrictive because the columns X_i of X can always be scaled to get (1.13). In this case, the entries of β^0 are scaled but neither $\mathbb{H}_0$ nor $\mathbb{H}_1$ are changed. Hence, without loss of generality, we admit to invoking an innocuous normalization on the columns of the design matrix. Remark also that (1.13) is satisfied under the stronger assumption, namely

$$\Theta = \mathrm{Id}_n \quad \text{and} \quad \forall i \in [p], \quad \|X_i\|_2^2 = 1. \tag{1.14}$$

Moreover, observe that, almost surely, there exists a unique couple $(\widehat{\imath}, \widehat{\varepsilon}) \in [p] \times \{\pm 1\}$ such that $\widehat{\varepsilon} U_{\widehat{\imath}} = \|U\|_\infty$. Under Assumption (1.13), the test statistic, referred to as *Spacing test for LARS*, simplifies to

$$S := \frac{\bar{\Phi}(\lambda_1)}{\bar{\Phi}(\lambda_2)}, \tag{1.15}$$

where we denote by Φ the cumulative distribution function of the standard normal distribution, $\bar{\Phi} = 1 - \Phi$ its complement, $\lambda_1 := \widehat{\varepsilon} U_{\widehat{\imath}}$ the largest knot in the *Lasso path* [EHJT04] and

$$\lambda_2 := \bigvee_{1 \le j \neq \widehat{\imath} \le p} \Big\{ \frac{U_j - R_{j\widehat{\imath}} U_{\widehat{\imath}}}{1 - \widehat{\varepsilon} R_{j\widehat{\imath}}} \vee \frac{-U_j + R_{j\widehat{\imath}} U_{\widehat{\imath}}}{1 + \widehat{\varepsilon} R_{j\widehat{\imath}}} \Big\},$$

with $a \vee b := \max(a, b)$ and U_i denotes the i-th entry of the vector U. Under Assumption (1.14), one has $R = X^\top X$ and λ_2 simplifies to the second largest knot in the *Lasso path*. Interestingly,

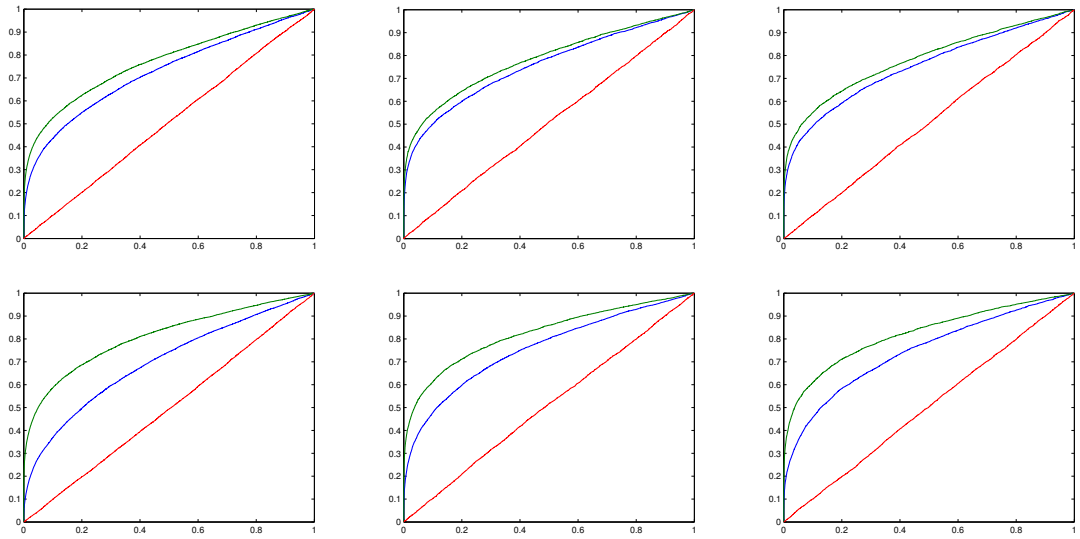

Figure 1.5: The spacing test and its "*studentization*" are exact.

the authors of [TLT13] have shown that the test statistic S is uniformly distributed on $[0,1]$ under the null hypothesis $\mathbb{H}_0$. Moreover, they derived the rejection region $\{S \leq \alpha\}$ for all $\alpha \in (0,1)$. In other words, the observed value of the test statistic S is the p-value of the spacing test. The next result shows that the spacing test is unbiased.

Theorem 9 ([DC10])**.** *Let $\alpha \in (0,1)$ be a significance level. Assume that the variance-covariance matrix Θ of the noise is known and assume that Assumption (1.13) holds. Then, the spacing test for LARS is unbiased: its power under the alternative is always greater or equal to the significance level α.*

Key step(s) of the proof: The proof is based on Anderson's inequality for symmetric convex sets and derives a simple and short proof of the distribution of the test statistic (1.15) under the null. ■

Under mild assumptions, this theorem ensures that the probability of getting a "*true positive*" is greater or equal to the probability of a "*false positive*". Moreover, in the limit case when the significance level α goes to zero, this result is refined in [DC10]: the probability of a true positive is much greater than the probability of getting a false positive. As a matter of fact, we prove that the cumulative distribution function of S has a vertical tangent at the origin under the alternative hypothesis. The reader may consult Figure 1.5 which represents the empirical distribution function of S that exactly describes the uniform law. Theorem 9 has a stronger version in the case of orthogonal designs, e.g. when the variance-covariance matrix Θ is Id_n and[14] $X^\top X = \mathrm{Id}_p$.

Theorem 10 ([DC10])**.** *Assume that $R = \mathrm{Id}_n$ then, under any alternative in $\mathbb{H}_1$, the density function of S is decreasing. Hence, for any significance level $\alpha \in (0,1)$, the region $\{S \leq \alpha\}$ is the most powerful region among all possible regions.*

Key step(s) of the proof: Thanks to a well chosen change of variables, we can compute exactly the distribution of the test statistic under the alternative. ■

[14] which implies that $n \geq p$.

Algorithm 1: t-Spacing test

Data: An observation $y \in \mathbb{R}^n$ and a design matrix $X \in \mathbb{R}^{n \times p}$.
Result: A p-value $T \in (0,1)$.

Compute the first LARS knot λ_1;

1. Set $U := X^\top y$;
2. Find $(\widehat{\imath}, \widehat{\varepsilon}) \in [p] \times \{\pm 1\}$ such that $\widehat{\varepsilon} U_{\widehat{\imath}} = \|U\|_\infty$ and set $\lambda_1 := \widehat{\varepsilon} U_{\widehat{\imath}}$;

Compute the second LARS knot λ_2;

3. Set $R := X^\top X$;
4. Set $\lambda_2 := \bigvee_{1 \le j \ne \widehat{\imath} \le p} \Big\{ \frac{U_j - R_{j\widehat{\imath}} U_{\widehat{\imath}}}{1 - \widehat{\varepsilon} R_{j\widehat{\imath}}} \vee \frac{-U_j + R_{j\widehat{\imath}} U_{\widehat{\imath}}}{1 + \widehat{\varepsilon} R_{j\widehat{\imath}}} \Big\}$,

Compute the variance estimator $\widehat{\sigma}$;

5. Set $R_{-\widehat{\imath}} := X_{-\widehat{\imath}}^\top (\mathrm{Id}_n - X_{\widehat{\imath}} X_{\widehat{\imath}}^\top) X_{-\widehat{\imath}}$;
6. Set $\widehat{\sigma} := \|R_{-\widehat{\imath}}^{-1/2} V_{-\widehat{\imath}}\|_2 / \sqrt{n-1}$ where

$$V_{-\widehat{\imath}} := (U_1 - R_{1\widehat{\imath}} U_{\widehat{\imath}}, \dots, U_{\widehat{\imath}-1} - R_{(\widehat{\imath}-1)\widehat{\imath}} U_{\widehat{\imath}}, U_{\widehat{\imath}+1} - R_{(\widehat{\imath}+1)\widehat{\imath}} U_{\widehat{\imath}}, \dots, U_p - R_{p\widehat{\imath}} U_{\widehat{\imath}});$$

Compute the p*-value* T;

8. Set $T_1 := \lambda_1 / \widehat{\sigma}$ and $T_2 := \lambda_2 / \widehat{\sigma}$;
9. Set $T := \frac{1 - \mathbb{F}_{n-1}(T_1)}{1 - \mathbb{F}_{n-1}(T_2)}$, where we denote by $\mathbb{F}_{n-1}$ the cumulative distribution function of the t-distribution with $n-1$ degree(s) of freedom.

1.3.3 Extension to unknown variance

Interestingly, we can derive from our analysis a *studentization* of the test statistic (1.15). Indeed, we consider the test statistic

$$T := \frac{1 - \mathbb{F}_{n-1}(T_1)}{1 - \mathbb{F}_{n-1}(T_2)},$$

where $\mathbb{F}_{n-1}$ denotes the cumulative distribution function of the t-distribution with $n-1$ degrees of freedom and T_1, T_2 are statistics that can be computed in cubic time (cost of one Singular Value Decomposition (SVD) of the design matrix) from the first knots of the LARS algorithm, see Algorithm 1. In the sequel, for each $i \in [p]$, we may denote by $X_{-i} \in \mathbb{R}^{n \times (p-1)}$ the sub-matrix of X where the i-th column X_i has been deleted and we may assume that it has rank n. Observe that this is a mild assumption in a high-dimensional context.

Theorem 11 (t-Spacing test for LARS **[DC10]**)**.** *Assume that the variance-covariance matrix* Θ *is* $\sigma^2 \mathrm{Id}_n$ *where* $\sigma > 0$ *is unknown and that for all* $i \ne j \in [p]$*, one has* $\|X_i\|_2 = 1$*,* $X_i \ne \pm X_j$ *and* X_{-i} *has rank* n*. Then, under the null* $\mathbb{H}_0$*, the statistic* T *described by Algorithm 1 is uniformly distributed on* $[0,1]$.

In particular, we derive a detection test of significance level α considering the rejection region $\{T \le \alpha\}$. One can empirically witness (see Figure 1.5 for instance) that the *t-Spacing test for LARS* is an interesting test statistic taking smaller values under the alternative hypothesis.

Indeed Figure 1.5 represents the empirical distribution function of 15,000 p-values coming from various scenarii. 5,000 p-values drawn under the null (red), 5,000 p-values of S under the alternative (green) and 5,000 p-values of T under the alternative (blue). At the top, the level of sparsity s is equal to 2. At the bottom, s is equal to 5. In both cases, from left to right, $(n,p)=(50,100)$, $(100,200)$ and $(100,500)$.

1.4 Prospects

1. An interesting perspective in Sensitivity Analysis is to find cheap computational methods that assess which entries are non significant. Thanks to the so-called "Randomized Pick-Freeze" (RPF) method developed in **[DC7]**, this concern can be stated in the frame of high dimensional Statistics and the various studies of the lasso estimator. In particular, it seems that this issue can be related to the "safe rule" methods [FGS15] which are screening rules that leverage the known sparsity (or an upper bound on it) of the solution by ignoring some variables during (or even before) the optimization process, hence speeding up solvers. This important topic is overlooked in Sensitivity Analysis and interesting results should be at hand using RPF point of view **[DC7]**.

2. A natural question for the RPF method concerns the estimation of interactions of higher orders (*i.e.,* Sobol indices of higher orders) which assess the "influence" on the output of the interaction of sets of some inputs. Yet the article **[DC7]** concerns with first order Sobol indices, one may extend the RPF method to higher order Sobol indices estimation. Interestingly, this issue might shed light on column wise correlated designs for which a specific analysis should be developed. In particular, standard arguments based on column/line/entry independence cannot be invoked here and it seems that "*ad hoc*" methods have to be found to deal with the inherent correlation structure of the design matrix involved in the RPF method. Simultaneously, standard proofs on support recovery results have to be recast to finely assess the discrepancy between indices of different orders of interactions.

3. In the article **[DC16]**, we present what should be the "*ideal*" rate function (in view of the tail of the Tracy-Widom law) that may lead to the "*right*" bounds on the RICs constants. A challenging issue would be to prove that this rate function holds for Gaussian and/or sub-Gaussian matrices, with independent entries say. Actually only the dependence of the rate function $\mathbb{W}(\rho,t)$ in the parameter ρ really matters when bounding the RIC constants. Furthermore, observe that we can afford a "sub-optimal" power of t in the rate function $t\mapsto\mathbb{W}(\rho,t)$ as shown by **[DC16]**.

4. The tools developed in the article **[DC10]** might be used to study the power of the spacing test under particular alternatives involving for instance a control of the "sparsity" and/or of the magnitude of the entries. In particular, it can be used to address some issues pointed to by some discussants of the article [LTTT14] such as the "power loss" effect. When only sparse models are to be considered, the main competitor of the spacing test will be some test based directly on the size of the largest estimated coefficient. As pointed by a referee of **[DC10]**, an interesting test statistic might be λ_1 or a scaled multiple of it when the variance σ has to be estimated. It would be very interesting to describe the alternatives for which the spacing test outperformed standard competitors such as the aforementioned testing procedure based on the size of the largest estimated coefficient.

5. One of the main challenge in Post-Selection Inference is the "non-Gaussianity" issue. Indeed, all the testing procedures are based on independence of orthogonal linear statistics through the so-called "Polyhedral lemma" [HTW15, Page 152]. This original characteristic is too restrictive when considering practical situations where Gaussian noise

cannot be assumed. A first result beyond this restriction might be found in the paper **[DC10]** where a studentized version of the testing procedure is assessed together with a new proof of the test significance. It may be interesting to consider weaker forms of noise, for instance log-concave noises, and to derive a spacing test in this frame. A related issue is to prove that the "*studentization*" of the spacing test **[DC10]** is unbiased.

Chapter 2

Extremal moment problem in Statistics

We have seen in Chapter 1 that a popular model in Statistics is the linear model when the sample size n may be as small as the "*sparsity*" times a logarithmic factor in the parameters dimension p. To get such result, one usually assumes that the design X satisfies a RIP-like property (see Section 1.2.2) and one considers an estimator $\widehat{\beta}$ solution to a convex program (see Section 1.2.1) that involves a regularizing norm (*e.g.,* the ℓ_1-norm). The combination of these two aspects, RIP-like design and ℓ_1-regularization, makes the estimator $\widehat{\beta}$ satisfy oracle inequalities guaranteeing that "*stable*" and "*robust*" recovery is possible, see Section 1.2.1.

In this chapter, we will work with deterministic designs and we will see that they satisfy none of the RIP-like properties. Hence we have lost one of the two pilars of high-dimensional Statistics and it seems that we cannot claim for the powerful results seen in the previous chapter. However, we can fruitfully capitalize on the ideas of ℓ_1-regularization to build estimators. When aiming at signed/complex Borel measures, this approach falls into the frame of the extremal moment problem, *i.e.,* finding the minimal total variation norm Borel measure with prescribed moments/Fourier coefficients.

2.1 Moments of signed measures

2.1.1 The truncated generalized moment problem

Denote by $(\mathbb{K}, d)$ a compact metric set and consider the Banach space $\mathbf{E} := (\mathscr{C}(\mathbb{K}, \mathbb{R}), \|\cdot\|_\infty)$ of real-valued continuous functions over $\mathbb{K}$ endowed with the supremum norm. Recall that its topological dual $\mathbf{E}^\star := (\mathscr{M}(\mathbb{K}, \mathbb{R}), \|\cdot\|_1)$ is the Banach space of real Borel measures endowed with the total variation norm $\|\cdot\|_1$ that can be defined as

$$\forall \mu \in \mathbf{E}^\star, \quad \|\mu\|_1 := \sup_{\|f\|_\infty \leq 1} \int_{\mathbb{K}} f \, \mathrm{d}\mu .$$

Consider $\mathbf{M} := (\varphi_0, \varphi_1, \ldots, \varphi_{n-1}) \in \mathbf{E}^n$ a "*Markov system*" defined by the following property

$$\forall k \in [n-1], \quad \forall a \in \mathbb{R}^n \setminus \{0\}, \quad \#\Big\{ t \in \mathbb{K} \quad \text{s.t.} \quad \sum_{j=0}^{k} a_j \varphi_j(t) = 0 \Big\} \leq k .$$

see [KN77, Pages 31-43] or **[DC1]** for further details. Markov systems are any family of continuous functions such that generalized polynomials of order k (*i.e.,* any nonzero linear combination $\sum_{j=0}^{k} a_j \varphi_j$) has at most k distinct roots.

Remark 2. On the domain $\mathbb{K} = [-1, 1]$, standard examples encompass algebraic moments ($\varphi_k = t^k$) and trigonometric moments ($\varphi_k = \cos(\pi k t/2)$). The Fourier basis ($\varphi_k = \exp(2\pi \imath k t)$ with $k = -f_c, \ldots, f_c$ and $n = 2f_c + 1$) on the one dimensional torus $\mathbb{K} = [0, 1)$ is a complex valued Markov system, this example will be treated in Section 2.2.

We begin with a solution to the standard truncated generalized moment problem. Define the cone of truncated moments as

$$\mathbf{C}_n := \Big\{ m \in \mathbb{R}^n \quad \text{s.t.} \quad \exists \mu \geq 0\,,\ m = \int_{\mathbb{K}} \mathbf{M}\,\mathrm{d}\mu \Big\},$$

where $\mu \geq 0$ denotes any positive measure $\mu \in \mathbf{E}^\star$. The truncated moment problem aims at characterizing this cone. Note that its dual cone is given by

$$\mathbf{C}_n^\star := \Big\{ a \in \mathbb{R}^n \quad \text{s.t.} \quad \forall t \in \mathbb{K},\ \sum_{j=0}^{n-1} a_j \varphi_j(t) \geq 0 \Big\}.$$

It follows that the truncated generalized moment problem is equivalent to characterizing nonnegative generalized polynomials of degree $n-1$. This result pertains to the "*full*" moment problem (*i.e.,* characterization of full sequences of algebraic moments) thanks to the Riesz-Haviland extension theorem, see [Las09, Theorem 3.1] for example.

In the case of algebraic moments, some important structure results exist. For instance, on a "*compact basic semi-algebraic set*[1]" the Putinar's Positivstellensatz shows that positive polynomials are Sum-of-Squares (SoS), see [Las09] for instance. Notice that SoS polynomials can be parametrized by semidefinite matrices, see for instance Chapters 2 in [Dum07, Las09]. Using this characterization, one can show that sequences of moments can be equivalently described using "*hierarchies*" of semidefinite matrices, see Section 2.1.3.

2.1.2 Beurling minimal extrapolation

In this section, we focus on the Exact Reconstruction property (as in Section 1.1.1) in the frame of finite Borel measure recovery. Precisely, given $m^0 \in \mathbb{R}^n$, consider the extremal generalized moment problem given by

$$\widehat{\mu} \in \arg \min_{m^0 = \int_{\mathbb{K}} \mathbf{M}\,\mathrm{d}\mu} \|\mu\|_1\,, \tag{2.1}$$

that aims at extending a finite number of moments m^0 by a representing measure $\widehat{\mu}$ of minimal total variation norm. This extremal moment problem has been intensively studied in various fields of Mathematics at the beginning of the 20th century[2]. The aforementioned estimator $\widehat{\mu}$ solution to (2.1) was recently studied in **[DC1]** (referred to as "*Beurling Minimal Extrapolation*", BME for short) and [CFG14] for instance. In these papers, the authors focus on the Fenchel dual program of (2.1) which reads as

$$\widehat{a} = \arg \min_{a \in \mathscr{P}_1(\mathbb{K},\mathbf{M})} \langle a, m^0 \rangle\,, \tag{2.2}$$

where we denote the dual feasible set by

$$\mathscr{P}_1(\mathbb{K},\mathbf{M}) := \Big\{ a \in \mathbb{R}^n \quad \text{s.t.} \quad \| \sum_{j=0}^{n-1} a_j \varphi_j \|_\infty \leq 1 \Big\}. \tag{2.3}$$

Recall that $\mathbf{M}$ is continuous over a compact set $\mathbb{K}$ and therefore uniformly bounded. We deduce that $\mathscr{P}_1(\mathbb{K},\mathbf{M})$ contains a small open ball around the origin. Generalized Slater conditions[3] show that there is no duality gap ("*strong duality*") and the Karush-Kuhn-Tucker (KKT) conditions give that the estimator $\widehat{\mu}$ satisfies implicit optimality equations

$$\|\widehat{\mu}\|_1 = \int_{\mathbb{K}} \widehat{P}\,\mathrm{d}\widehat{\mu} \quad \text{where} \quad \widehat{P} := \sum_{j=0}^{n-1} \widehat{a}_j \varphi_j \in \mathscr{P}_1(\mathbb{K},\mathbf{M}) \quad \text{and} \quad m^0 = \int_{\mathbb{K}} \mathbf{M}\,\mathrm{d}\widehat{\mu}\,. \tag{2.4}$$

[1] see (2.7) for a definition.

[2] Arne Beurling [Beu38] initiated the theory of extension functions in Harmonic Analysis when studying the minimal total variation norm function among all bounded variation functions with prescribed Fourier transform on a given domain.

[3] see [BC11, Proposition 26.18] for instance.

Conversely any $\widehat{\mu}$ satisfying (2.4) is a solution to (2.1). A first step toward faithful recovery is given by the following result where we denote by δ_t the Dirac mass at point $t \in \mathbb{K}$. Also, by an abuse of notation, we may denote $P = \sum_{j=0}^{n-1} a_j \varphi_j \in \mathscr{P}_1(\mathbb{K}, \mathbf{M})$ for $a \in \mathscr{P}_1(\mathbb{K}, \mathbf{M})$.

Theorem 12 ([DC1]). *If there exists a polynomial $P^0 \in \mathscr{P}_1(\mathbb{K}, \mathbf{M})$, points $t_1^0, \ldots, t_s^0 \in \mathbb{K}$ and signs $\varepsilon_1^0, \ldots, \varepsilon_s^0 \in \{\pm 1\}$ such that $P^0(t_\ell^0) = \varepsilon_\ell^0$ for $\ell \in [s]$ and $|P(t)| < 1$ for $t \neq t_1^0, \ldots, t_s^0$ then, for all $\mu^0 = \sum_{\ell=1}^s \alpha_\ell^0 \delta_{t_\ell^0} \in \mathbf{E}^\star$ such that sign of α_ℓ^0 equals ε_ℓ^0, it holds that μ^0 is the unique solution to (2.1) when setting $m^0 = \int_{\mathbb{K}} \mathbf{M} \, \mathrm{d}\mu^0$.*

Key step(s) of the proof: By optimality and by construction of P^0, one has

$$\|\mu^0\|_1 = \int_{\mathbb{K}} P^0 \mathrm{d}\mu^0 = \int_{\mathbb{K}} P^0 \mathrm{d}\widehat{\mu} \leq \|\widehat{\mu}\|_1 \leq \|\mu^0\|_1 ,$$

and, in particular, $\int_{\mathbb{K}} P^0 \mathrm{d}\widehat{\mu} = \|\widehat{\mu}\|_1$ which shows that $\widehat{\mu}$ is supported by $\{t_1^0, \ldots, t_s^0\}$. The Markov system property gives that measures with common support $\{t_1^0, \ldots, t_s^0\}$ are identifiable, leading to $\widehat{\mu} = \mu^0$. ■

The polynomial P^0 is called the "*dual certificate*", it guarantees faithful recovery of μ^0 given the observation of some moments m^0 whenever the support of μ^0 is included in $t_1^0, \ldots, t_s^0 \in \mathbb{K}$ and the corresponding weights have signs $\varepsilon_1^0, \ldots, \varepsilon_s^0 \in \{\pm 1\}$. An example is given in Figure 2.1.

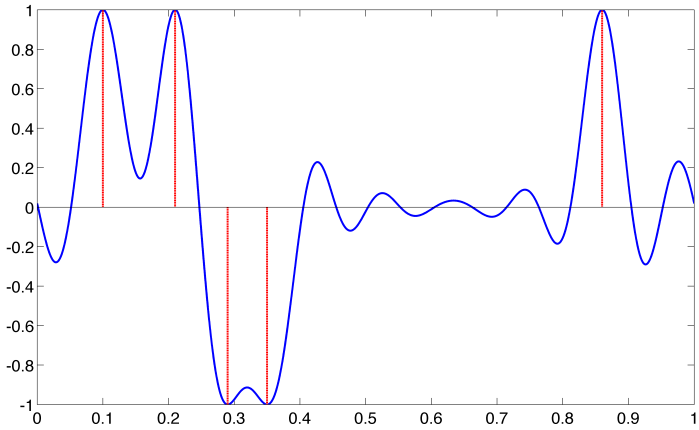

Figure 2.1: A dual certificate P^0.

Theorem 13 ([DC1]). *If $\mathbf{M}$ is a Markov system, $n \geq 2s+1$ and $\varepsilon_1^0 = \cdots = \varepsilon_s^0$ then the corresponding dual certificate P^0 exists.*

Key step(s) of the proof: A standard property of Markov systems [KN77] shows that one can build a nonnegative generalized polynomial Q of degree $n-1$ such that $Q(t_\ell^0) = 0$ and $Q(t) > 0$ for $t \neq t_1^0, \ldots, t_s^0$. Denote by $\varepsilon := \varepsilon_1^0 = \cdots = \varepsilon_s^0$ and set $P^0 = \varepsilon(1 - tQ)$ for $t > 0$ sufficiently small. ■

This theorem shows that in the special case where the target measure $\mu^0 \in \mathbf{E}^\star$ is atomic and its amplitudes α_ℓ share the same sign/phase then it is possible to faithfully recover μ^0 minimizing the total variation norm.

Theorem 14 ([DC1]). *Let $\mathbf{M}$ be a Markov system and let μ^0 be a measure with finite support. If the sample size n exceeds a bound that depends exponentially on the inverse of the minimum distance $\min_{i \neq j} d(t_i^0, t_j^0)$ then μ^0 is the unique solution to (2.1) when setting $m^0 = \int_{\mathbb{K}} \mathbf{M} \, \mathrm{d}\mu^0$.*

Key step(s) of the proof: We give the proof for the algebraic polynomial case since the same construction can be derived for general Markov systems, see [KN77] for useful properties on those systems of functions.

Denote by Δ the minimum distance $\min_{i\neq j} d(t_i^0, t_j^0)$. Consider the Lagrange interpolation polynomials $L_k(t) := \prod_{\ell\neq k}(t-t_\ell^0)/\prod_{\ell\neq k}(t_k^0-t_\ell^0)$ for $k\in[s]$. One can upper bound the supremum norm of L_k over $\mathbb{K}=[0,1]$ by a bound $L(\Delta)$ that depends only on the minimum distance Δ. Consider the m-th Chebyshev polynomial of the first kind defined for all $x\in[-1,1]$ by $T_m(x) := \cos(m\arccos(x))$. For a sufficiently large value of m, there exists s extrema x_ℓ of T_m such that $|x_\ell| \leq 1/(sL(\Delta))$ and $T_m(x_\ell) = \varepsilon_\ell$. Interpolating values x_ℓ at points t_ℓ, we build the dual certificate as

$$P^0(t) = T_m\Big(\sum_{\ell=1}^{s} x_\ell L_\ell(t)\Big).$$

We find that the polynomial P^0 has a degree no greater than $2/\sqrt{\pi}(\sqrt{e}/\Delta)^{5/2+1/\Delta}$. Note that this polynomial is such that $P^0(t_\ell^0) = \varepsilon_\ell^0$ but it does not satisfy $|P(t)|<1$ for $t\neq t_1^0,\ldots,t_s^0$ though $\|P^0\|_\infty \leq 1$. However, using a Vandermonde matrix (derived from the property of Markov systems), we conclude the proof. ■

This result shows that if the support points $t_1^0,\ldots,t_s^0\in\mathbb{K}$ are sufficiently separated then there exists a dual certificate whatever the signs/phase of the amplitudes α_ℓ. This result has been recently improved in the Fourier case by the paper [CFG14], see Section 2.2.1.

2.1.3 Lasserre's hierarchies and Sum-of-Squares

Primal program (2.1) can be efficiently solved using Lasserre's hierarchies **[DC11]** or from the dual program (2.2) using Sum-of-Squares (SoS) representation [CFG14].

∘ We present a construction of a solution to (2.1) using the dual program (2.2) and SoS representation of the constraint $\mathscr{P}_1(\mathbb{K},\mathbf{M})$, see (2.3) for a definition. Consider the algebraic, trigonometric or Fourier univariate case, see Remark 2. Observe that

$$\mathscr{P}_1(\mathbb{K},\mathbf{M}) = \Big\{a\in\mathbb{R}^n \quad \text{s.t.} \quad \forall t\in\mathbb{K},\ \underbrace{1-|\sum_{j=0}^{n-1} a_j\varphi_j(t)|^2}_{Q(t)} \geq 0\Big\}$$

Note that the univariate polynomial Q can be decomposed[4] onto the basis $\varphi_0,\varphi_1,\ldots,\varphi_{2n-1}$. It follows [Dum07, CFG14] that representing $\mathscr{P}_1(\mathbb{K},\mathbf{M})$ reduces to representing nonnegative polynomials. Then the Fejér-Riesz theorem shows that univariate nonnegative polynomials are SoS polynomials[5]. Furthermore, one knows ([Dum07, Section 3.9.3] or [Las09, Page 17]) that SoS polynomials can be represented using Linear Matrix Inequality (LMI), *i.e.,* using a semidefinite constraint given by

$$\text{LMI constraint} := \Big\{z\in\mathbb{R}^N \quad \text{s.t.} \quad \mathbf{S}_0 + \sum_{j=1}^{N} z_j\mathbf{S}_j \succeq 0\Big\} \tag{2.5}$$

where $\succeq 0$ means positive semidefinite matrix, $N\geq 0$ and $\mathbf{S}_0,\mathbf{S}_1,\ldots,\mathbf{S}_N$ are symmetric matrices. It results [CFG14] that the dual program (2.2) is a semidefinite program (SDP) and it can be solved using an interior point method for sizes up to $n=\mathscr{O}(1e5)$ on standard 2016 laptops.

Once the dual is solved, we extract the dual solution $\widehat{P}$ and, in view of (2.4), it holds

$$\widehat{S} := \operatorname{supp}(\widehat{\mu}) \subseteq \{t\in\mathbb{K} \quad \text{s.t.} \quad |\widehat{P}(t)|=1\}, \tag{2.6}$$

where $\operatorname{supp}(\widehat{\mu})$ denotes the support of the finite Borel measure $\widehat{\mu}\in\mathbf{E}^\star$. We deduce [CFG14] that if the polynomial $|\widehat{P}|^2$ is not constant then the support $\widehat{S}$ of $\widehat{\mu}$ is finite and included in the roots of the derivative of the polynomial $|\widehat{P}|^2$. This procedure is referred to as the "*root-finding*" step. Then uncovering $\widehat{\mu}$ reduces to solve a linear system of equations $m^0 = \int_{\mathbb{K}} \mathbf{M}\,\mathrm{d}\widehat{\mu}$ knowing the support $\widehat{S}$. This can be easily done since the system is invertible by property of the Markov-systems.

[4] In a semidefinite manner with respect to (a_j) using Schur complement, see [Dum07, Eq. (4.34)].

[5] Actually they are squares of a single univariate polynomial.

∘ We can circumvent some previous method limitations with the construction of a solution to (2.1) using the primal program and Lasserre's hierarchies, see **[DC11]** for further details. This setting concerns with the trigonometric polynomial basis and the algebraic polynomial basis in dimension $d \geq 1$. For sake of readability, we focus on the algebraic polynomial basis given by $t^k := t_1^{k_1} \cdots t_d^{k_d}$ and we assume that $\mathbf{M}$ is given by the family t^k where $k \in \mathbf{K}$ for some finite $\mathbf{K} \subset \mathbb{N}^d$. Consider I polynomials $g_1, \ldots, g_I$ and assume that

$$\mathbb{K} = \Big\{ t \in \mathbb{R}^d \quad \text{s.t.} \quad \forall i \in I,\ g_i(t) \geq 0 \Big\} \tag{2.7}$$

is compact with an "*algebraic certificate of compactness*[6]". The set defined by (2.7) is referred to as a compact basic semi-algebraic set[7]. Reminding Jordan decomposition of any signed measure $\mu = \mu^+ - \mu^-$ where $\mu^\pm \geq 0$, consider the identity

$$\begin{array}{lll} \min & \|\mu\|_1 & = \quad \min \quad \Big\{ m_0^+ + m_0^- \Big\} \\ \text{s.t.} & m^0 = \int_{\mathbb{K}} \mathbf{M}\, \mathrm{d}\mu & \text{s.t.} \quad m_k^+ - m_k^- = m_k^0, \quad \text{for } k \in \mathbf{K}, \\ & & \quad m^+ \in \mathbf{C}(\mathbb{K}) \\ & & \quad m^- \in \mathbf{C}(\mathbb{K}) \end{array} \tag{2.8}$$

where $m^0 = (m_k^0)_{k \in \mathbf{K}}$ and the infinite-dimensional moment cone is given by

$$\mathbf{C}(\mathbb{K}) := \Big\{ (m_k)_{k \in \mathbb{N}^d} \quad \text{s.t.} \quad \forall k \in \mathbb{N}^d,\ m_k = \int_{\mathbb{K}} t^k\, \mu(\mathrm{d}t),\ \mu \geq 0 \Big\}.$$

Using Putinar's Positivstellensatz [Las09, Page 29] and the duality argument presented in Section 2.1, one can prove that

$$\mathbf{C}(\mathbb{K}) = \bigcap_{r \geq 0} \underbrace{\Big\{ (m_k)_{k \in \mathbb{N}^d} \quad \text{s.t.} \quad \mathscr{M}_r(m_k) \succeq 0 \Big\}}_{\mathbf{C}^r(\mathbb{K})} \tag{2.9}$$

where $\mathscr{M}_r(m_k) \succeq 0$ is a LMI (2.5) on a finite number of $(m_k)_{k \in \mathbb{N}^d}$ with known symmetric matrices $\mathbf{S}_j$ depending on $g_1, \ldots, g_I$. More precisely, the matrix $\mathscr{M}_r(m_k)$ is a $I+1$ blocks diagonal matrix and each square block has dimension $\binom{d+r-\lfloor \deg g_i/2 \rfloor}{d}$ and correspond to a "*localizing*" matrix, see [Las09, Page 61]. Furthermore, the sequence $(\mathbf{C}^r(\mathbb{K}))_{r \geq 0}$ in (2.9) is nested and, replacing $\mathbf{C}(\mathbb{K})$ by $\mathbf{C}^r(\mathbb{K})$ for a fixed $r \geq 0$, we get a semidefinite relaxation of (2.8).

The key remark is that the full and truncated moment cones are difficult to represent and the strategy developed here consists in building outer nested semidefinite representable approximations $\mathbf{C}^r(\mathbb{K}) \supseteq \mathbf{C}^{r+1}(\mathbb{K}) \supseteq \mathbf{C}^{r+2}(\mathbb{K}) \supseteq \cdots \supseteq \mathbf{C}(\mathbb{K})$ whose limit infimum is $\mathbf{C}(\mathbb{K})$. This sequence of semidefinite relaxation is referred to as "*Lasserre's hierarchies*". In general, one can prove that the solutions given by a subsequence of relaxations converge toward the target. An interesting feature is to prove that the solution of the relaxation of finite order r is exactly the target solution to the primal. In practice this situation can be certified using a "*rank stabilization*" argument [Las09, Chapter 4] and, in this case, one can extract a finite support measure representing the moment sequence. From a theoretical point of view, some finite convergence results can be proved [Nie14] when the primal objective function is linear and random.

In **[DC11]**, we present this methodology in the frame of the extremal moment problem. We proved that the hierarchy converges to the true minimum as r goes to infinity. We showed that one can detect finite convergence of the hierarchy using "*rank stabilization*" of the localizing matrices and, in this case, one can extract a finite support solution $\widehat{\mu}$. Of course, finite convergence of the hierarchy occurs for trigonometric polynomials on $\mathbb{K} = [0,1)$, which follows from the Fejér-Riesz theorem and this was exploited in the landmark paper [CFG14]. Similarly, but apparently not so well-known, a weaker form of the Fejér-Riesz theorem also holds[8] for

[6] Such certificate can be enforced adding the polynomial $R - \|t\|_2^2$ to the g_i's, with $R > 0$ sufficiently large.

[7] In particular, a finite union of shifted ℓ_p-balls for different $p \in [1, \infty]$ is a compact basic semi-algebraic set.

[8] Indeed it follows from [Sch06, Corollary 3.4] that every nonnegative bivariate trigonometric polynomial can be written as a sum of squares of trigonometric polynomials.

bivariate trigonometric polynomials. So again for trigonometric polynomials on $\mathbb{K} = [0,1)^2$, finite convergence of the hierarchy takes place. Note however that in contrast to the one-dimensional case, there is no explicit upper bound on the degrees of the sum of squares (SoS) which are required, so that even in the two-dimensional Fourier case we do not have an a priori estimate of the smallest value of r for which rank stabilization occurs.

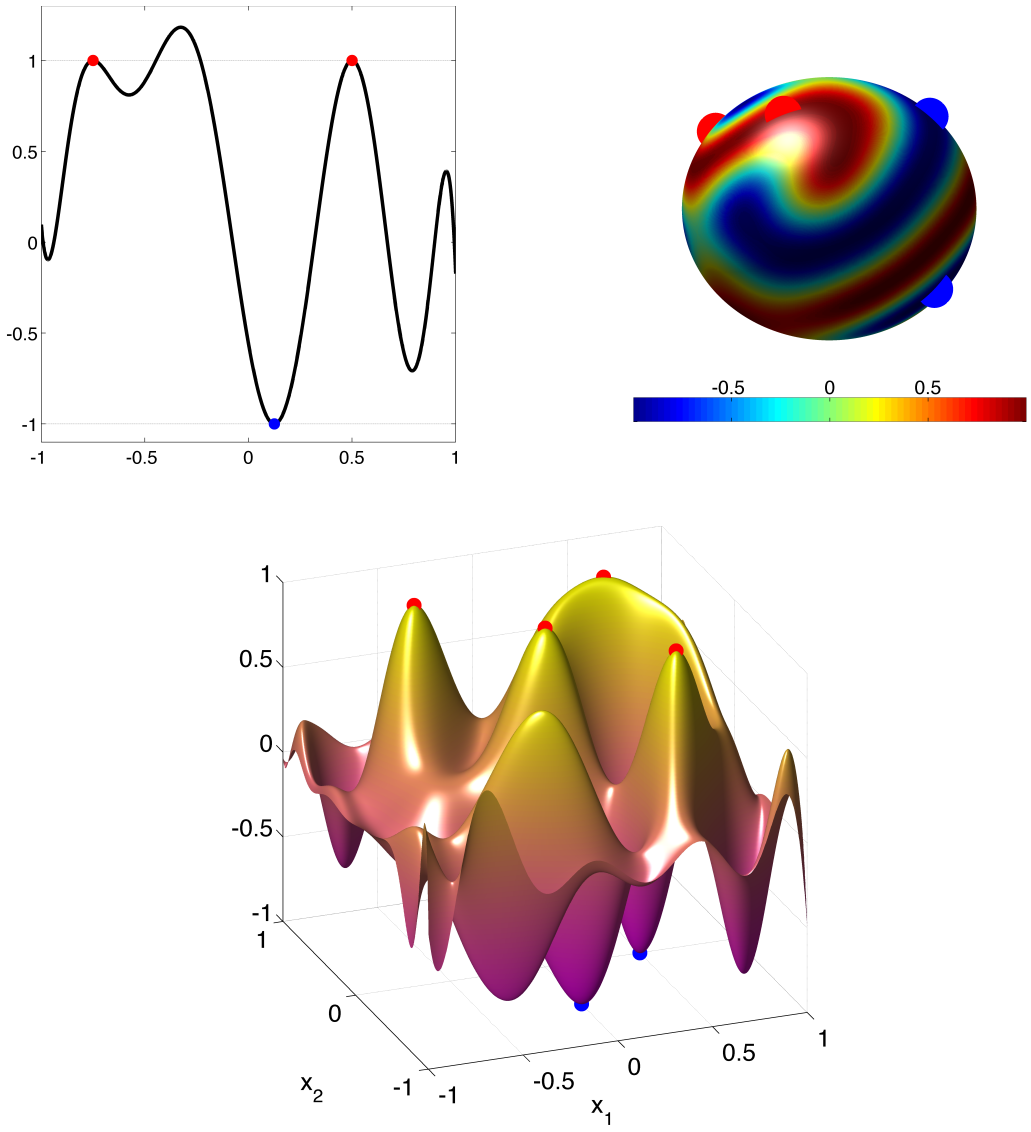

Figure 2.2: Lasserre's hierarchies solutions on various domains.

In Figure 2.2 we have presented three examples of **[DC11]**. On the top left, a degree 9 polynomial certificate for the univariate example $\mathbb{K} = [-1,-1/2] \cup [0,1]$, with 2 points (red) in the support of the positive part, and 1 point (blue) in the support of the negative part of the optimal measure. On the top right, a degree 6 polynomial certificate for the sphere example, with 3 points (red) in the support of the positive part, and 3 points (blue) in the support of the negative part of the optimal measure. On the bottom, a degree 12 polynomial certificate for the bivariate example $\mathbb{K} = [-1,1]^2$, with 4 points (red) in the support of the positive part, and 2 points (blue) in the support of the negative part of the optimal measure.

2.2 Super-Resolution

The statistical analysis of the ℓ_1-regularization in the space of measures was initiated by Donoho [Don92] and then investigated by [DG96, GG96]. Recently, this problem has attracted a lot of attention in the "*Super-Resolution*" community and its companion formulation in "*Line spectral estimation*". In the Super-Resolution frame, one aims at recovering fine scale details of an image from few low frequency measurements, ideally the observation is given by a low-pass filter. The novelty of this body of work lies in new theoretical guarantees of the ℓ_1-minimization over the space of discrete measures in a gridless manner, as presented in Section 2.1. Some recent work on this topic can be found in [BP13, TBSR13, CFG14, CFG13, FG13, BDF16, DP15a], **[DC1]**, **[DC5]**, **[DC6]**, **[DC12]**. More precisely, pioneering work can be found in [BP13] which treats of inverse problems on the space of Radon measures and [CFG14] which investigates the Super-Resolution problem via Semi-Definite Programming and the ground breaking construction of a "*dual certificate*", see Section 2.2.1. Exact Reconstruction property (in the noiseless case), minimax prediction and localization (in the noisy case) have been performed using the "*Beurling Lasso*" estimator (2.11) introduced in **[DC5]** and also studied in [TBSR13, FG13, TBR15] which minimizes the total variation norm over complex Borel measures, see Program (2.11). Noise robustness (as the noise level tends to zero) has been investigated in the absorbing paper [DP15a], the reader may also consult [DP15b, DDP15]. Change point detection and grid-less spline decomposition are studied in [BDF14] and **[DC6]**. Several interesting extensions (such as the deconvolution over spheres) are shown in [BDF15a, BDF15b, BDF16].

2.2.1 The separation condition

One can prove **[DC12]** that standard conditions on design matrices such as the Restricted Isometry Property (RIP), Restricted Eigenvalue Condition (REC), or the Compatibility Condition cannot hold in the present frame. But, one can also prove **[DC1]** that the notion of null space property (NSP) of Section 1.1.2 can be extended to the frame of Borel measure recovery when the sampling scheme is given by a Markov system. More precisely, the key for proving NSP is based on a general construction of dual certificates for Markov systems. The paper **[DC1]** proved that if the sampling size n exceeds a bound that depends exponentially on the inverse of the distance $d(t_i^0, t_j^0)$ then the dual certificate P^0 of Theorem 12 exists, see Theorem 14. This bound can be dramatically improved in the Fourier case, as shown by the ground-breaking paper [CFG14]. The emanating condition is referred to as the "*space out*" condition by **[DC1]** or the "*separation*" condition by [CFG14], see Assumption 1.

Now, we introduce the framework of Super-Resolution, *i.e.,* reconstruction of a complex Borel measure $\mu^0 \in \mathbf{E}^\star$ on the one dimensional[9] torus $\mathbb{K} = [0,1)$ from the observation of some Fourier coefficients. Precisely, our observation vector is $y \in \mathbb{C}^n$ where $n = 2f_c + 1$ and the integer $f_c \geq 1$ is referred to as the "*frequency cut off*" of the low-pass filter. Our sampling scheme is modeled by the linear operator $\mathscr{F}_n$ that maps a complex Borel measure to its first Fourier coefficients, namely

$$\forall \mu \in \mathbf{E}^\star, \quad \mathscr{F}_n(\mu) := (m_k(\mu))_{|k| \leq f_c} \quad \text{where} \quad m_k(\mu) := \int_{\mathbb{T}} \overline{\varphi_k} \mathrm{d}\mu,$$

where we consider the Fourier basis on the torus $\mathbb{T} := [0,1)$ given by

$$\varphi_k(t) := \exp(2\pi \imath k t),$$

for $k \in \{-f_c, \ldots, 0, \ldots, f_c\}$, throughout this section.

[9] The one dimension case is the most studied and, up to technicalities, the results can be extended to higher dimensions by a tensorization argument.

The model we consider is formulated as

$$y = \mathscr{F}_n(\mu^0) + \zeta \tag{2.10}$$

with ζ a complex valued centered Gaussian random variable defined by $\zeta = \zeta^{(1)} + \imath\zeta^{(2)}$ where the real part $\zeta^{(1)} = \Re(\zeta)$ and the imaginary part $\zeta^{(2)} = \Im(\zeta)$ are i.i.d. $\mathcal{N}_n(0, \sigma_0^2 \mathrm{Id}_n)$ random vectors with standard deviation $\sigma_0 > 0$. Moreover, we assume that the target measure μ^0 admits a sparse structure, namely it has finite support and reads

$$\mu^0 = \sum_{\ell=1}^{s} \alpha_\ell^0 \delta_{t_\ell^0},$$

where $s \geq 1$, δ_t is the Dirac mass at point $t \in \mathbb{T}$ and $\alpha_\ell^0 \in \mathbb{C}$. We can now state the "*separation condition*".

Assumption 1 (Separation condition)**.** The target support $\mathrm{supp}(\mu^0) = \{t_1^0, \dots, t_s^0\}$ verifies the separation condition if

$$\forall i, j \in [s], \quad \text{s.t.} \quad i \neq j, \qquad d(t_i^0, t_j^0) \geq \frac{c_0}{f_c},$$

where $d(\cdot,\cdot)$ is the wrap-around distance on the torus $\mathbb{T} = [0,1)$ and $c_0 := 1.26$ and $f_c \geq 10^3$.

Under this assumption, the papers [CFG14, FG16] show that the dual certificate P^0 of the target μ^0 exists and thus proves the Exact Recovery property, see Theorem 12. Interestingly, the construction of [CFG14, FG16] is sufficiently versatile to be extended to other ℓ_∞-constrained interpolation problems, see [TBR15] or **[DC12]** for instance. These constructions have originated the minimax prediction and localization results presented in the next section.

2.2.2 Minimax prediction and localization

In view of the analysis of ℓ_1-regularization presented in Chapter 1, the paper **[DC5]** introduces the "*Beurling Lasso*" (Blasso) as

$$\widehat{\mu} \in \arg\min_{\mu \in \mathbf{E}^*} \left\{ \frac{1}{2} \|y - \mathscr{F}_n(\mu)\|_n^2 + \lambda \|\mu\|_1 \right\}, \tag{2.11}$$

where $\lambda > 0$ is a tuning parameter and we recall that $\|\cdot\|_n := \|\cdot\|_2/\sqrt{n}$. One can prove that the prediction $\mathscr{F}_n(\widehat{\mu})$ of (2.11) is unique. Then, using the results of Section 2.1, one may find an atomic representing measure $\widehat{\mu}$ with minimal total variation norm, see Section 2.2.4 for further details. Hence we may consider an atomic solution $\widehat{\mu} = \sum_{\ell=1}^{\widehat{s}} \widehat{\alpha}_\ell \delta_{\widehat{t}_\ell}$ in the sequel.

Following **[DC5]** and [CFG14, FG16], we define the set of "*near*" points of the target support $\mathrm{supp}(\mu^0) = \{t_1^0, \dots, t_s^0\}$ as

$$\forall \ell \in [s], \quad N_\ell := \left\{ t \in \mathbb{T}; \quad d(t, t_\ell^0) \leq \frac{c_1}{f_c} \right\},$$

for some $c_1 \in (0, c_0/2)$ (where c_0 as given in Assumption 1) and the set of "*far*" points as

$$F := \mathbb{T} \setminus \bigcup_{\ell \in [s]} N_\ell .$$

This partition allows the characterization of the localization performances of BLasso as shown in the next theorem.

Theorem 15 (**[DC5]** and [TBR15, FG16])**.** *There exist universal constants $\gamma, C, C' > 0$ such that the following holds. Under Assumption 1, the estimator $\widehat{\mu}$ solution to Problem (2.11) with a choice $\lambda \geq C\sigma_0\sqrt{\log n/n}$ satisfies, $\forall \ell \in [s]$,*

$$\left| a_\ell - \sum_{\{k:\, \widehat{t}_k \in N_\ell\}} \widehat{a}_k \right| \leq C' s\lambda, \qquad \sum_{\{k:\, \widehat{t}_k \in N_\ell\}} |\widehat{a}_k| d^2(t_\ell^0, \widehat{t}_k) \leq C' \frac{s\lambda}{n^2}, \quad \text{and} \quad \sum_{\{k:\, \widehat{t}_k \in F\}} |\widehat{a}_k| \leq C' s\lambda$$

with probability at least $1 - C' n^{-\gamma}$.

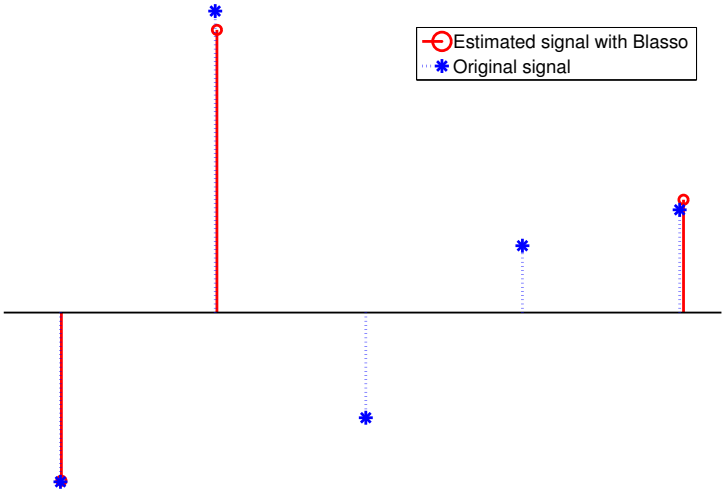

Figure 2.3: Localization properties of the Beurling Lasso.

Key step(s) of the proof: This proof can be found in **[DC5]**, **[DC6]** and **[DC12]**. It begins with the following control of the supremum of the Gaussian process

$$\forall u>0,\quad \mathbb{P}\Big\{\|n^{-1}\sum_{|k|\leq f_c}\zeta_k\exp(2\pi\imath kt)\|_\infty>u\Big\}\leq n\exp\Big(-n\frac{u^2}{4\sigma_0^2}\Big),$$

using a Kac-Rice formula [AW09, Pages 69-71]. In high-dimensional Statistics the aforementioned quantity is simply $\lambda_0:=n^{-1}\|X^\top\zeta\|_\infty$ and a standard optimization argument shows that λ has to be chosen greater than this quantity. It leads to the relation $\lambda\geq C\sigma_0\sqrt{\log n/n}$. Then we prove a fine control of the Bregman divergence of the total variation norm at the target point μ^0, namely

$$0\leq D_{\mathrm{TV}}(\widehat{\mu},\mu^0):=\|\widehat{\mu}\|_{\mathrm{TV}}-\|\mu^0\|_{\mathrm{TV}}-\mathscr{R}\left(\int_{\mathbb{T}}\overline{P^0}(t)(\widehat{\mu}-\mu^0)(\mathrm{d}t)\right)\leq Cs\lambda.$$

where P^0 is the dual certificate associated to the target μ^0. Using other interpolating polynomials (given by a modification of the construction of [CFG14]) and some algebraic manipulations, one concludes the proof. ■

As displayed in Figure 2.3, typical properties of the Blasso are that it estimates well large spikes, it can miss small spikes[10] and it can put small spikes in the "far" region, *i.e.*, it overestimates the support (small false positives occur in the "far" region of the target support) as in the Lasso case. In particular, we deduce the following result that supports the idea that large spikes (*i.e.*, such that $|a_\ell^0|\gg s\lambda$) are well estimated.

Corollary 2 (**[DC5]** and [TBR15, FG16])**.** *Let $C>2\sqrt{2}$. There exist universal constants $\gamma,C'>0$ such that the following holds. Under Assumption 1, the estimator $\widehat{\mu}$ solution to Problem (2.11) with a choice of tuning parameter $\lambda\geq C\sigma_0\sqrt{\log n/n}$ satisfies that, for any t_ℓ^0 in the support of μ^0 verifying $|a_\ell^0|>C's\lambda$, there exists an element $\widehat{t}_k$ in the support of $\widehat{\mu}$ such that*

$$\mathrm{d}(t_\ell^0,\widehat{t}_k)\leq\sqrt{\frac{C's\lambda}{|a_\ell^0|-C's\lambda}\frac{1}{n}}$$

with probability at least $1-C'n^{-\gamma}$.

[10] Which can be viewed as a soft-thresholding effect as in the Lasso case.

As for the prediction problem, the Super-Resolution setting meets with the curse of highly correlated designs since close Dirac masses share almost the same Fourier coefficients. In particular, one can prove **[DC12]** that standard conditions, such as the Compatibility Condition [vdGB09, vdG16] or RIP [CT05, CT06], cannot hold in the present frame. Yet, one can prove [TBR15] and **[DC12]** that BLasso prediction performances achieve a "*fast rate*" of convergence[11], namely $\sigma_0 s/n$ up to a log factor.

Theorem 16 ([TBR15] and **[DC12]**)**.** *Let $C > 2\sqrt{2}$. There exists universal constants $\gamma, C' > 0$ such that the following holds. Assume Assumptions 1 and 2. The estimator $\widehat{\mu}$ solution to Problem (2.11) with a choice of tuning parameter $\lambda \geq C\sigma_0\sqrt{\log n/n}$ satisfies*

$$\|\mathscr{F}_n(\widehat{\mu}-\mu^0)\|_n^2 \leq C' s\,\lambda^2,$$

with probability at least $1-C'n^{-\gamma}$.

In **[DC12]**, this result is formulated for a slightly different estimator, the Square-root Blasso ($\sqrt{B}$Lasso), see Section 2.2.3. In particular, both estimators achieve the minimax rate[12] in prediction up to a log factor.

2.2.3 Simultaneous noise and signal estimation

Note that the estimation of the noise level in sparse high dimension regression appeared first in [Owe07, Ant10], as well as in [XCM10] with a game theory flavor. Later on, in [SBvdG10], the authors proposed to analyze another variant that is a convex reformulation of a penalized joint log-likelihood estimator. An equivalent definition of this estimator was proposed and extensively studied independently in [BCW11] under the name "*Square-root Lasso*". We adopt this terminology for the estimator (2.12), though our formulation is inspired by the one proposed in [SZ12] under the name "*Scaled-Lasso*". Our proposed contribution **[DC12]** borrows some ideas from the stimulating lecture [vdG16]. Hence, we introduced the "*Square-root Blasso*" ($\sqrt{B}$Lasso) estimator that jointly estimates the signal and the noise level

$$(\widehat{\mu},\widehat{\sigma}) \in \operatorname*{argmin}_{(\mu,\sigma)\in\mathbf{E}^*\times\mathbb{R}_{++}} \Big\{\frac{1}{2\sigma}\|y-\mathscr{F}_n(\mu)\|_n^2 + \frac{\sigma}{2} + \lambda\|\mu\|_{\mathrm{TV}}\Big\}. \tag{2.12}$$

where $\mathbb{R}_{++}$ denotes the set of positive real numbers. This new estimator can be efficiently computed using Fenchel-Legendre duality and a semi-definite representation of nonnegative trigonometric polynomials, see Section 2.2.4.

A standard assumption [vdG16] in the analysis of the Square-root Lasso governs the Signal-to-Noise Ratio (SNR) which can be defined as $\mathrm{SNR} := \|\mu^0\|_{\mathrm{TV}}/(\sqrt{2}\sigma_0)$ measuring the strength of the true signal μ^0 compared to the noise level $\sqrt{\mathbb{E}\|\zeta\|_n^2} = \sqrt{2}\sigma_0$.

Assumption 2 (Sampling rate condition)**.** The sampling rate condition holds if and only if

$$\lambda\cdot\mathrm{SNR} \leq (\sqrt{17}-4)/2 \simeq 0.0616\,.$$

Remark 3. The main point here is to consider a noise-free tuning parameter λ that depends only on the sample size n. We consider $\lambda \geq C\sqrt{\log n/n}$ where $C > 2\sqrt{2}$ is some numerical constant. In this regime, one may write the sampling rate condition as $n/\log n \geq C'\mathrm{SNR}^2$ for some universal constant $C' > 0$. Roughly speaking, Assumption 2 states that the number of measurements n is at least SNR^2.

Another important assumption is the "*no-over fitting*" condition [vdG16, Section 3.2] assuming that the noise level estimator $\widehat{\sigma}$ is positive. For obvious reasons, it is essential from both theoretical and practical points of view to assert this property. This is done by the following proposition.

[11] see [DHL15] for interesting results on rates of convergence of prediction performances for Lasso.
[12] In [TBR15], the minimax rate is derived using the minimax rate of [CP09] for sparse regression.

Proposition 17 ([vdG16] and **[DC12]**). *Let $\eta \in (0,1)$ and $\alpha \in (0,1)$. Recall that λ is the tuning parameter of the $\sqrt{B}$Lasso (2.12). Set $\underline{\sigma} := \sqrt{2}\sigma_0\big(1-\sqrt{-2\log\alpha/n}\big)^{1/2}$ and $\lambda_0 := \sqrt{2\log(n/\alpha)/n}$. If $\lambda \geq (1-\eta)^{-1}\lambda_0$ and*

$$\lambda \frac{\|\mu^0\|_{\mathrm{TV}}}{\underline{\sigma}} \leq 2\Big(\sqrt{1+(\eta/2)^2}-1\Big), \tag{2.13}$$

then it holds that $|\widehat{\sigma}/\|\zeta\|_n - 1| \leq \eta$ with probability larger than $1-\alpha\Big(\frac{2\sqrt{2}}{n}+\frac{2\sqrt{3}+3}{3}\Big)$.

Key step(s) of the proof: Set $\mathscr{F}_n^*\zeta := \sum_{|k|\leq f_c} \zeta_k \exp(2\pi\imath k t)$. Using a Kac-Rice formula on a non-Gaussian process [AW09, Page 79], we get that

$$\forall u>0, \quad \mathbb{P}\Big\{\frac{n^{-1}\|\mathscr{F}_n^*\zeta\|_\infty}{\|\zeta\|_n} > u\Big\} \leq \Big(2\sqrt{2}+\frac{2n}{\sqrt{3}}\Big)\Big(1-\frac{u^2}{2}\Big)^n.$$

A standard concentration argument (see [BLM13] for instance) shows that $\mathbb{P}\{\|\zeta\|_n \leq \underline{\sigma}\} \leq \alpha$. A union bound on the event $\{n^{-1}\|\mathscr{F}_n^*\zeta\|_\infty/\|\zeta\|_n \leq R\} \cap \{\|\varepsilon\|_n \geq \underline{\sigma}\}$ combined with standard argument on the "no-over fitting" condition [vdG16, Lemma 3.1] gives the result. ∎

Note that Assumption 2 implies (2.13) with $\eta = 1/2$. Choosing $\alpha = o(n)$ and $\eta > 0$ arbitrarily small, Proposition 17 shows that $\widehat{\sigma}/\sqrt{2}$ is a consistent[13] estimator of σ_0.

2.2.4 Dual polynomials

We have seen three estimators, Beurling Minimal Extrapolation (BME) described in (2.1), Beurling Lasso (2.11) and Square-root Blasso (2.12). Aside from Lasserre's hierarchies that have been presented in Section 2.1.3, we have seen that a standard technic to compute the estimator $\widehat{\mu}$ is to solve the dual program that estimates the coefficients of a trigonometric polynomial $\widehat{P}$ referred to as the "*dual polynomial*". By the Karush–Kuhn–Tucker (KKT) condition, the support of the estimated measure $\widehat{\mu}$ is included in the roots of the derivative of the polynomial $|\widehat{P}|^2$, as we have seen at (2.6) in Section 2.1. In Table 2.1, we present the primal and dual formulations of those estimators where we denote

$$\mathscr{P}_1(\mathbb{T},\mathscr{F}_n) := \Big\{a \in \mathbb{R}^n \quad \text{s.t.} \quad \forall t \in \mathbb{T},\ 1-\Big|\sum_{k=-f_c}^{f_c} a_j \varphi_k(t)\Big|^2 \geq 0\Big\},$$

$$\mathscr{P}_1'(\mathbb{T},\mathscr{F}_n) := \Big\{a \in \mathbb{R}^n \quad \text{s.t.} \quad \forall t \in \mathbb{T},\ 1-\Big|\sum_{k=-f_c}^{f_c} a_j \varphi_k(t)\Big|^2 \geq 0 \quad \text{and} \quad n\lambda^2\|a\|_2^2 \leq 1\Big\},$$

the dual feasible sets.

Estimator	Primal	Dual
BME	$\min_{m^0=\mathscr{F}_n\mu} \|\mu\|_1$	$\max_{a\in\mathscr{P}_1(\mathbb{T},\mathscr{F}_n)} \langle m^0, a\rangle$
BLasso	$\min_{\mu\in\mathbf{E}^\star} \Big\{\frac{1}{2}\|y-\mathscr{F}_n(\mu)\|_n^2 + \lambda\|\mu\|_1\Big\}$	$\max_{a\in\mathscr{P}_1(\mathbb{T},\mathscr{F}_n)} \frac{1}{n}\langle y,a\rangle - \frac{\lambda}{2}\|a\|_2^2$
$\sqrt{B}$Lasso	$\min_{(\mu,\sigma)\in\mathbf{E}^\star\times\mathbb{R}_{++}} \Big\{\frac{1}{2\sigma}\|y-\mathscr{F}_n(\mu)\|_n^2 + \frac{\sigma}{2} + \lambda\|\mu\|_{\mathrm{TV}}\Big\}$	$\max_{a\in\mathscr{P}_1'(\mathbb{T},\mathscr{F}_n)} \lambda\langle y,a\rangle$

Table 2.1: Super-Resolution estimators and their duals.

[13] Recall that $\mathbb{E}\|\zeta\|_n^2 = 2\sigma_0^2$.

In the BME and Blasso cases, the knowledge of the support of $\widehat{\mu}$ suffices to recover the primal solution inverting a Vandermonde system, see (2.6). However, for the $\sqrt{B}$Lasso case, the variance estimator $\widehat{\sigma}$ has to be found also and we cannot recover the measure estimator $\widehat{\mu}$ from its support $\{\widehat{t}_1,\dots,\widehat{t}_{\widehat{s}}\}$ and the KKT conditions as before. This case can be solved considering the estimators $(\widehat{a},\widehat{\sigma})$ given by

$$(\widehat{a},\widehat{\sigma}) \in \underset{(\alpha,\sigma)\in\mathbb{R}^{\widehat{s}}\times\mathbb{R}_{++}}{\operatorname{argmin}} \frac{1}{2\sigma}\|y-X\alpha\|_n^2 + \frac{\sigma}{2} + \lambda\|\alpha\|_1, \tag{2.14}$$

where the design matrix $X \in \mathbb{C}^{n\times\widehat{s}}$ is defined by $X_{k,j} = \overline{\varphi_k}(\widehat{t}_j)$. One can check that $(\widehat{\mu},\widehat{\sigma})$ satisfies the original optimality condition for problem (2.12), where we recall that $\widehat{\mu} = \sum_{j=1}^{\widehat{s}} \widehat{a}_j \delta_{\widehat{t}_j}$. To solve (2.14), we proceed following the alternate minimization procedure [SZ12], that consists in alternating between a Lasso step and a noise level $\widehat{\sigma}$ estimation step[14]. Note that the Lasso step is elementary in this case, since the KKT conditions lead to $\widehat{a} = X^\dagger y - \lambda n\widehat{\sigma}(X^*X)^{-1}\widehat{\varepsilon}$ where $X^\dagger$ denotes the Moore-Penrose pseudoinverse and $\widehat{\varepsilon}$ are the phases of the dual polynomial $\widehat{P}$ at the estimated points $\widehat{t}_\ell$. This method is introduced in **[DC12]**.

An interesting feature **[DC12]** of the $\sqrt{B}$Lasso (2.12) is that its dual solutions $\widehat{P}$ are never constant, so the "*root-finding*" step (seen at (2.6) in Section 2.1) can be always invoked.

2.3 Experimental designs

It occurs to the authors of **[DC11]** that solving the "*approximate optimal experimental design problem*" can be recast as a discrete measure reconstruction under moments constraints. Hence, all the techniques that have been developed in this chapter can be invoked in this framework. Although **[DC18]** is being finalized, I choose to present here some of its results since they may open an interesting field in my (future) work.

2.3.1 Convex design theory

The optimum experimental designs are computational and theoretical objects that minimize the variance of the best linear unbiased estimators in regression problems. In this frame, the experimenter models response y_i of a random experiment whose input parameters are represented by a vector $t_i \in \mathbb{R}^d$ with respect to known regression functions $\mathbf{M} := (\varphi_1,\dots,\varphi_p)$, namely for all $i \in [N]$, one has $y_i = \sum_{j=1}^p \theta_j \varphi_j(t_i) + \varepsilon_i$ where $\theta \in \mathbb{R}^p$ are unknown parameters that the experimenter wants to estimate, ε_i is some noise and t_i is chosen by the experimenter in a *design space* $\mathbb{K} \subset \mathbb{R}^d$. Assume that the distinct points among $t_1,\dots,t_N$ are the points $t_1,\dots,t_s$, for some $s \in [N]$, and let N_i denote the number of times the particular point t_i occurs among $t_1,\dots,t_N$, for all $i \in [s]$. This would be summarized by

$$\zeta := \begin{pmatrix} t_1 & \cdots & t_s \\ \frac{N_1}{N} & \dots & \frac{N_s}{N} \end{pmatrix}, \tag{2.15}$$

whose first row gives the points in the *design space* $\mathbb{K}$ where the inputs parameters have to be taken and the second row tells the experimenter which proportion of experiments ("*frequencies*") have to be done at these points. The goal of the design of experiment theory is then to assess which inputs parameters t_i and frequencies $w_i := N_i/N$ the experimenter has to consider. For a given ζ, the standard analysis of the Gaussian linear model shows that the minimal covariance matrix (with respect to Loewner ordering) of unbiased estimators can be expressed in terms of the Moore-Penrose pseudoinverse of the *information matrix* which is defined by

$$\mathbb{M}(\zeta) := \sum_{i=1}^s w_i \mathbf{M}(t_i)\mathbf{M}^\top(t_i). \tag{2.16}$$

[14] Consisting in computing the norm of the residual.

As a matter of fact, one major aspect of design of experiment theory seeks to maximize the information matrix over the set of all possible ζ. Notice the Loewner ordering $\succcurlyeq$ is partially ordered and, in general, there is no greatest element among all possible matrices $\mathbb{M}(\zeta)$. The standard approach is to consider some statistical criteria, namely the *Kiefer's ϕ_p-criteria* [Kie74], in order to describe and construct the "*optimum designs*" with respect to those criteria. Observe that the *information matrix* $\mathbb{M}(\zeta)$ belongs to $\mathbb{S}_p^+$, the space of symmetric nonnegative definite matrices of size p, and define, for all $q \in [-\infty, 1]$, a criterion ϕ_q where for positive definite matrices M it holds

$$\phi_q(M) := \begin{cases} (\frac{1}{p}\mathrm{trace}(M^q))^{1/q} & \text{if } q \neq -\infty, 0 \\ \det(M)^{1/p} & \text{if } q = 0 \\ \lambda_{\min}(M) & \text{if } q = -\infty \end{cases}$$

and for nonnegative definite matrices M it reads $\phi_q(M) := (\frac{1}{p}\mathrm{trace}(M^q))^{1/q}$ if $q \in (0, 1]$, and zero otherwise. Those criteria are meant to be real valued, positively homogeneous, non constant, upper semi-continuous, isotonic (with respect to the Loewner ordering $\succcurlyeq$) and concave functions. In particular, we search for solutions to the following optimization problems

$$\zeta^\star \in \arg \max_{\zeta \text{ as in (2.15)}} \phi_q(\mathbb{M}(\zeta)), \tag{2.17}$$

where the maximum is taken over all design matrices ζ of the form (2.15) and $q \in [-\infty, 1]$.

Observe that the set of admissible designs described by (2.15) is any combination of s pairwise distinct support points t_i in the *design space* $\mathbb{K}$ and number of replications N_i at t_i such that $\sum_i N_i = N$. It appears that the set of admissible frequencies $w_i = N_i/N$ is discrete and contained in the set of rational numbers of the form a/N where a is an integer. Hence, notice that (2.17) is a discrete optimization problem with respect to frequencies w_i. To the best our of knowledge, this combinatorial problem is extremely difficult both analytically and computationally. A popular solution is then to consider "*approximate*" designs defined by

$$\zeta := \begin{pmatrix} t_1 & \cdots & t_s \\ w_1 & \cdots & w_s \end{pmatrix}, \tag{2.18}$$

where w_i are varying continuously from 0 to 1 and $\sum_{i=1}^s w_i = 1$. Accordingly, any solution to (2.17) where the maximum is taken over all matrices of type (2.18) is called "*approximate optimal design*".

2.3.2 A Linear Matrix Inequality formulation

We assume again that $\mathbb{K}$ is a compact semi-algebraic set[15] with an algebraic certificate of compactness. Moreover, we assume that $\mathbf{M} \subset \mathbb{R}_n[x]^p$ where $\varphi_\ell(t) := \sum_{k \in \{0,\ldots,n\}^d} \mathbf{a}_{\ell,k} t^k$. Notice that these assumptions cover a large class of problems in optimal design theory, see for instance [DS97, Chapter 5]. Define, for all $\mu \geq 0$, the information matrix (with an abuse of notation)

$$\mathbb{M}(\mu) = \Big(\int_{\mathbb{K}} \varphi_i \varphi_j \mathrm{d}\mu \Big)_{1 \leq i,j \leq p} = \Big(\sum_{k,t \in \{0,\ldots,d\}^n} \mathbf{a}_{i,k} \mathbf{a}_{j,t} m_{k+t}(\mu) \Big)_{1 \leq i,j \leq p}.$$

Note that $\mathbb{M}(\mu) = \sum_{|\alpha| \leq 2d} m_\alpha(\mu) \mathbf{A}_\alpha$ where for all $\alpha \in \{0,\ldots,2d\}^m$,

$$\mathbf{A}_\alpha := \Big(\sum_{k+\ell=\alpha} \mathbf{a}_{i,k} \mathbf{a}_{j,\ell} \Big)_{i,j}.$$

Further, set $\mu = \sum_{i=1}^\ell w_i \delta_{t_i}$ and observe that $\mathbb{M}(\mu) = \sum_{i=1}^\ell w_i \mathbf{M}(t_i) \mathbf{M}^\top(t_i)$ as in (2.16). Recall that the ϕ_q-criteria for $q \in [-\infty, 1]$ are isotonic with respect to the Loewner ordering $\succcurlyeq$ and

[15] Remind the definition at (2.7).

then, for all $X \in \mathbb{S}_p^+$ and for all $\mu \in \mathbf{E}^\star$,

$$\Big\{ \sum_{\alpha\in\{0,\dots,2d\}^m} m_\alpha(\mu)\mathbf{A}_\alpha - X \succcurlyeq 0 \Big\} \Rightarrow \Big\{ \phi_q(\mathbb{M}(\mu)) \geq \phi_q(X) \Big\} \tag{2.19}$$

We deduce the following Linear Matrix Inequality (LMI) equivalent formulation of our problem

$$\zeta^\star \in \arg\max_{X\in\mathscr{D}_0(\mathbb{K},\mathbf{M})} \phi_q(X), \tag{2.20}$$

where the feasible set $\mathscr{D}_0(\mathbb{K},\mathbf{M})$ is given by

$$\mathscr{D}_0(\mathbb{K},\mathbf{M}) := \Big\{ X \in \mathbb{S}_p^+ \,:\, \sum_{|\alpha|\leq 2d} m_\alpha(\mu)\mathbf{A}_\alpha - X \succcurlyeq 0,\ \mu = \sum_{i=1}^s w_i \delta_{t_i} \geq 0,\ \sum_i^s w_i = 1,\ s \geq 1 \Big\},$$

and designs ζ can be identified with atomic probabilities μ. In particular, note that $\zeta^\star$ is identified to $\mu^\star$ such that $X^\star = \sum_{\alpha\in\{0,\dots,2d\}^m} m_\alpha(\mu^\star)\mathbf{A}_\alpha$, since that, by isotonicity, the constraint (2.19) is active at the solution point $X^\star$ of (2.20) .

2.3.3 Solving the approximate optimal design problem

Let us introduce a two step procedure to solve (2.20). The first step focuses on a characterization of the truncated moment cone $\mathscr{M}_{2d} = \Big\{ (m_\alpha(\mu))_{\alpha\in\{0,\dots,2d\}^m} \,:\, \mu \geq 0,\ m_0(\mu) = 1 \Big\}$. Note that, by the Carathéodory theorem, the truncated moment cone is exactly

$$\mathscr{M}_{2d} := \{ (m_\alpha(\mu))_{\alpha\in\{0,\dots,2d\}^m} \,:\, \mu = \sum_{i=1}^s w_i \delta_{t_i} \geq 0,\ \sum_i^s w_i = 1,\ s \geq 1 \}.$$

So that we consider $(m_\alpha^\star)_{\alpha\in\{0,\dots,2d\}^m}$ a solution to

$$(m_\alpha^\star)_{\alpha\in\{0,\dots,2d\}^m} \in \arg\max_{X\in\mathscr{D}_1(\mathbb{K},\mathbf{M})} \phi_q(X) \tag{2.21}$$

where the feasible set $\mathscr{D}_1(\mathbb{K},\mathbf{M})$ is given by

$$\mathscr{D}_1(\mathbb{K},\mathbf{M}) := \Big\{ X \in \mathbb{S}_p^+ \,:\, \sum_{|\alpha|\leq 2d} m_\alpha \mathbf{A}_\alpha - X \succcurlyeq 0,\ (m_\alpha)_{\alpha\in\{0,\dots,2d\}^m} \in \mathscr{M}_{2d} \Big\},$$

and we identify $(m_\alpha^\star)_{\alpha\in\{0,\dots,2d\}^m}$ thanks to the active constraint $X^\star = \sum_{\alpha\in\{0,\dots,2d\}^m} m_\alpha^\star \mathbf{A}_\alpha$. Interestingly, the truncated moment cone $\mathscr{M}_{2d}$ can be represented using Lasserre's hierarchies. It follows that (2.21) can be efficiently solved using those hierarchies, see **[DC18]**. In practice, we may witness finite convergence of the hierarchies[16] so that the solution of the SDP relaxation is exactly the solution to (2.21).

Once we have exactly solved step one, we have to find a representing atomic measure $\mu^\star$ of the truncated moment sequence $(m_\alpha^\star)_{\alpha\in\{0,\dots,2d\}^m}$. This problem can be also worked out using Lasserre's hierarchies. Indeed, invoking Jiawang Nie's trick [Nie14] (*i.e.,* minimizing any random linear functional), we can prove that finite convergence of Lasserre's hierarchies almost surely occurs and "*extract*" the solution $\zeta^\star$.

2.4 Prospects

1. An important open question [CFG14] is to assess when the dual polynomial is constant (in absolute value). Writing the descent cone of the dual program at points corresponding to constant polynomials, it seems that this event can be related to the Gaussian measure of some moment cone. Indeed, one may see that the descent cone is a rotated version of the positive moment cone. An interesting perspective is to precisely bound this probability in the Super-Resolution frame and/or in the general frame of Markov systems.

[16] For instance observing rank stabilization of the solution matrices.

2. We have started a work **[DC19]** on post selection inference in Super Resolution. This framework is far from being exhausted and it seems to be an interesting prospect for future work. In particular, it would be appealing to find a testing procedure from the Blasso or the $\sqrt{B}$Lasso solutions for fixed λ. The "polyhedral lemma" (see for instance [HTW15, Page 152]) cannot be invoked here since the conditioning event is no longer "polyhedral" and it is a negligible Borel set. We understand that a new analysis has to be developed in this context together with new testing procedures.

3. It would be interesting to study the use of Lasserre's hierachies for the $\sqrt{B}$Lasso and/or the Blasso. An alternating direction method might help solve the primal formulation (2.12) for instance. Indeed, one can consider a two-step procedure updating the estimated variance and then solving the Blasso with the updated variance estimate. It seems that a new proof has to be found proving finite convergence of the hierarchies in this context.

4. In [FG16], several interesting extensions (point sources with a common support or demixing of sines and spikes) of Super-Resolution are presented though with no theoretical study. In particular one can pursue two goals when treating those cases : estimating the noise level thanks to the Rice method and searching for prediction and localization error bounds. These questions seems to be valuable topics in Super-Resolution though there is no theoretical analysis yet.

5. Poisson noise models have not been investigated yet in the Super-Resolution frame. This model corresponds to a photon count, where the noise intensity is proportional to the number of photons that hit the receptor during the exposition time. This is useful to model medical imaging, tomography imaging and digital camera noises. A Poisson noise model would require a new analysis and new tools when deriving prediction and localization error bounds.

6. Proving finite convergence of Lasserre's hierarchies in Program (2.8), it would be interesting to study the estimator given by replacing the objective function $\{m_0^+ + m_0^-\}$ by $\{w_+ m_0^+ + w_- m_0^-\}$ in (2.8) where w_+, w_- are random positive weights. Indeed, using Jiawang Nie's trick [Nie14] (*i.e.,* minimizing any random linear functional), we can prove that finite convergence of Lasserre's hierarchies almost surely occurs. Furthermore, it seems that this adjustment will not affect the Exact Recovery property using the appropriate notion of dual certificate.

Chapter 3

Latent space models

Latent space models are widely used statistical models for which the law of the observations depends on a hidden structure that is possibly random as well. The goal is then to recover the latent structure and the (conditional) distribution of the observations. This general model encompasses an extremely broad variety of theoretical and practical situations and the purpose of this chapter is not meant to present the plentiful topics emanating from latent space models but rather aggregate some contributions on nonparametric hidden Markov models and graph reconstruction from eigen spaces.

3.1 Nonparametric hidden Markov models

Finite state space hidden Markov models (HMMs for short) are widely used to model data evolving in time and coming from heterogeneous populations. They seem to be reliable tools to model practical situations in a broad class of fields such as economics, genomics, signal processing and image analysis, ecology, environment, speech recognition, to name but a few.

3.1.1 Model, assumptions and identifiability

From a statistical view point, finite state space HMMs are stochastic processes $(X_n, Y_n)_{n\geq 1}$ where $(X_n)_{n\geq 1}$ is a Markov chain with finite state space $[X]$ for $X \geq 1$ and transition matrix $\mathbf{Q}^\star$. The key feature is that conditionally on $(X_n)_{n\geq 1}$ the Y_n's are independent with a distribution depending only on X_n, namely $\mathscr{L}((Y_n)_{n\geq 1}|(X_n)_{n\geq 1}) = \bigotimes_{n\geq 1} \mathscr{L}(Y_n|X_n)$. Given a sample chain size $N \geq 3$, the observations are $Y_{1:N} = (Y_1, \ldots, Y_N)$ and the associated states $X_{1:N} = (X_1, \ldots, X_N)$ are unobserved. The parameters of the model are the initial distribution $\pi^\star$, the transition matrix of the hidden chain $\mathbf{Q}^\star$, and the "*emission distributions*" of the observations, that is the probability distributions of the Y_k's conditionally to $\{X_k = x\}$ for all possible states $x \in [X]$. Assume that there exists a compact $\mathscr{Y} \subset \mathbb{R}^d$ such that $Y_1 \in \mathscr{Y}$ almost surely and, conditionally to $\{X_k = x\}$, it has a density with respect to the Lebesgue measure $\mathscr{L}^d$ on $\mathscr{Y} \subset \mathbb{R}^d$. To be specific, we define the "*emission densities*" as

$$\forall x \in [X], \quad f_x^\star := \frac{\mathrm{d}\mathscr{L}(Y_1|X_1 = x)}{\mathrm{d}\mathscr{L}^d}.$$

The preliminary obstacle to obtaining theoretical results on general finite state space nonparametric HMMs was to understand when such models are indeed identifiable. Marginal distributions of finitely many observations were finite mixtures of products of the emission distributions. It is clear that identifiability cannot be obtained based on the marginal distribution of only one observation. The papers [AMR09, HKZ12, AHK12] paved the way to obtaining identifiability under reasonable assumptions. In [AHK12] the authors point out a structural link between multivariate mixtures with conditionally independent observations and finite state space HMMs. In [HKZ12] the authors give a spectral method to estimate all parameters for finite state space HMMs (with finitely many observations), under the assumption that the

transition matrix $\mathbf{Q}^\star$ of the hidden chain is non singular, and that the (finitely valued) emission distributions are linearly independent. Those spectral methods are particularly interesting since they do not suffer from initialization issues as in standard methods in latent model estimation such as the Expectation Maximization algorithm, see Section 3.1.3.

Extension to emission distributions on any space, under the linear independence assumptions (and keeping the assumption of non singularity of the transition matrix), allowed for the proof of the general identifiability result for finite state space HMMs. Indeed [GCR16] proved that if the emission densities are linearly independent and the transition matrix has full rank then the transition matrix $\mathbf{Q}^\star$ and the emission densities $f_1^\star, \dots, f_X^\star$ are identifiable from the distribution of three consecutive observations (Y_1, Y_2, Y_3) up to label switching of the hidden states. One fundamental result originating in the first works [AMR09] on identifiability is due to J. Kruskal [Kru77] who proved that, under certain explicit conditions, the expression of a third-order tensor[1] of rank r as a sum of r tensors of rank one is unique, up to permutation of the summands. Later, [AH16] obtained a non constructive proof of identifiability when the emission distributions are all distinct (not necessarily linearly independent) and still when the transition matrix of the hidden chain is full rank. In [GCR16], model selection likelihood methods and nonparametric kernel methods are also proposed to get nonparametric estimators. Minimax adaptive estimation in nonparametric HMM is proved in **[DC8]** (see Section 3.1.2) and nonparametric estimation of the filtering and marginal smoothing distributions in **[DC14]** (see Section 3.1.5). Standard assumptions on the hidden chain and emission laws are listed below.

Assumption (H)

- The transition matrix $\mathbf{Q}^\star$ has full rank,
- The Markov chain $(X_n)_{n\geq 1}$ is irreducible and aperiodic,
- The initial distribution $\pi^\star = (\pi_1^\star, \dots, \pi_X^\star)$ is the stationary distribution,
- The family of emission densities $\mathfrak{F}^\star := \{f_1^\star, \dots, f_X^\star\}$ is linearly independent.

These assumptions appear in spectral methods, see for instance [HKZ12, AHK12], in identifiability issues, see for instance [AMR09, GCR16] and in nonparametric estimation, filtering and smoothing, see **[DC8]** and **[DC14]**.

3.1.2 Penalized least-squares method

We begin with some notation. We assume that the emission densities $(f_x^\star)_{x\in[X]}$ belong to $(\mathbf{L}^2(\mathscr{Y}, \mathscr{L}^d), \|\cdot\|_2)$ the Hilbert space of square integrable functions on $\mathscr{Y}$. Consider $(\mathfrak{P}_M)_{M\geq 1}$ a nested sequence of subspaces of dimension M such that their union is dense in $\mathbf{L}^2(\mathscr{Y}, \mathscr{L}^d)$. The sequence $(\mathfrak{P}_M)_{M\geq 1}$ defines a family of sieves that we use as approximation spaces to estimate the emission densities. Let $\Phi_M := \{\varphi_1, \dots, \varphi_M\}$ be an orthonormal basis of $\mathfrak{P}_M$, for instance splines, Fourier basis or wavelets, see **[DC8]**. Define the projection of the emission laws onto the subspaces $\mathfrak{P}_M$ by

$$f_{M,x}^\star := \sum_{m=1}^{M} \langle f_x^\star, \varphi_m \rangle \varphi_m \,, \tag{3.1}$$

for all states $x \in [X]$. We write $f_M^\star := (f_{M,1}^\star, \dots, f_{M,X}^\star)$ and $f^\star := (f_1^\star, \dots, f_X^\star)$. Further, for any $f = (f_1, \dots, f_X) \in (\mathbf{L}^2(\mathscr{Y}, \mathscr{L}^d))^X$ and any transition matrix $\mathbf{Q}$, denote by $g^{\mathbf{Q},f}$ the function

$$g^{\mathbf{Q},f}\left(y_1, y_2, y_3\right) = \sum_{x_1,x_2,x_3=1}^{X} \pi(x_1)\mathbf{Q}(x_1, x_2)\mathbf{Q}(x_2, x_3) f_{x_1}(y_1) f_{x_2}(y_2) f_{x_3}(y_3), \tag{3.2}$$

[1] *i.e.*, a 3-way array.

where π is the stationary distribution of $\mathbf{Q}$. When $\mathbf{Q} = \mathbf{Q}^\star$ and $f = f^\star$, we get $g^{\mathbf{Q}^\star, f^\star} = g^\star$. When $f = (f_1, \dots, f_X)$ are probability densities on $\mathscr{Y}$, $g^{\mathbf{Q},f}$ is the probability distribution of three consecutive observations of a stationary HMM. Consider the following assumption on the emission distribution.

Assumption (F)

Let $\mathscr{F}$ be a closed bounded set of $(\mathbf{L}^2(\mathscr{Y}, \mathscr{L}^d), \|\cdot\|_2)$ and $(\mathbf{L}^\infty(\mathscr{Y}, \mathscr{L}^d), \|\cdot\|_\infty)$ invariant by projection onto the $\mathfrak{P}_M$'s. Assume that $f^\star \in \mathscr{F}^X$, *i.e.,* the target emission densities belong to $\mathscr{F}$.

Denote by $\mathscr{Q}_X$ the space of irreducible matrices of size $X \times X$. For $\mathbf{Q} \in \mathscr{Q}_X$, consider the model $\mathscr{S}_{\mathbf{Q},M}$ of distributions of three consecutive observations when the transition matrix of the hidden Markov chain is $\mathbf{Q}$. Specifically, define by

$$\mathscr{S}_{\mathbf{Q},M} := \left\{ g^{\mathbf{Q},f} \quad \text{s.t.} \quad f \in (\mathscr{F} \cap \mathfrak{P}_M)^X \right\},$$

a collection of models indexed by their complexity $M \geq 1$. Furthermore, we assume an estimator $\widehat{\mathbf{Q}}$ of $\mathbf{Q}^\star$, for instance the spectral estimator (see Section 3.1.3), and we look at an estimator $\widehat{g}$ of $g^\star$ among the models $(\mathscr{S}_{\widehat{\mathbf{Q}},M})_{M\geq 1}$. To this end, we use the so-called "*penalized least-squares method*" as follows.

The least squares adjustment is made on the density $g^\star$ of (Y_1, Y_2, Y_3). Starting from the operator $\Gamma : \ t \mapsto \|t - g^\star\|_2^2 - \|g^\star\|_2^2 = \|t\|_2^2 - 2\int t g^\star$ which is minimal for the target $g^\star$, we introduce the corresponding empirical contrast γ_N,

$$\forall t \in \mathbf{L}^2(\mathscr{Y}^3, (\mathscr{L}^d)^{\otimes 3}), \quad \gamma_N(t) := \|t\|_2^2 - \frac{2}{N}\sum_{s=1}^{N-2} t(Z_s),$$

with $Z_s := (Y_s, Y_{s+1}, Y_{s+2})$. As the sample size N tends to infinity, $\gamma_N(t) - \gamma_N(g^\star)$ converges almost surely to $\|t - g^\star\|_2^2$, thus the name least squares contrast function. A natural estimator is then a function t such that $\gamma_N(t)$ is minimal over a judicious collection of models, for instance $(\mathscr{S}_{\widehat{\mathbf{Q}},M})_{M\geq 1}$ following **[DC8]**. We define a whole collection of estimates $(\widehat{g}_M)_{M\geq 1}$, each M indexing the approximation subspace $\mathscr{S}_{\widehat{\mathbf{Q}},M}$ by

$$\widehat{g}_M = \arg\min_{t \in \mathscr{S}_{\widehat{\mathbf{Q}},M}} \gamma_N(t). \tag{3.3}$$

It then remains to select a suitable model, that is to choose the M which minimizes the criterion $\|\widehat{g}_M - g^\star\|_2^2 - \|g^\star\|_2^2$. This quantity is close to $\gamma_N(\widehat{g}_M)$, but we need to take into account the deviations of the process $\Gamma - \gamma_N$. We rather minimize

$$\widehat{M} = \arg\min_{M=1,\dots,N} \left\{ \gamma_N(\widehat{g}_M) + \mathrm{pen}(N, M) \right\},$$

where $\mathrm{pen}(N, M)$ is a penalty term to be specified. Then the estimator of $g^\star$ is $\widehat{g} = \widehat{g}_{\widehat{M}}$, and the estimator of $f^\star$ is $\widehat{f} := \widehat{f}_{\widehat{M}}$ so that $\widehat{g} = g^{\widehat{\mathbf{Q}},\widehat{f}}$.

The penalized least-squares estimator does not have an explicit form such as in usual nonparametric estimation, so that one has to use numerical algorithms to minimize the empirical contrast γ_N over models $\mathscr{S}_{\widehat{\mathbf{Q}},M}$. As initial point of the minimization algorithm, we shall use the spectral estimator, see Section 3.1.4 for more details. Since the spectral estimator is consistent (see Theorem 21), the algorithm does not suffer from initialization problems in practice. From a theoretical point of view, one can derive an oracle inequality for the estimation of $g^\star$ using the model selection theory. Denote by $\mathfrak{S}_X$ the set of permutations τ on $[X]$.

Theorem 18 (**[DC8]**). *Assume **(H)** and **(F)**. Then there exist positive constants N_0, $\rho^\star$, C such that, if $\mathrm{pen}(N, M) \geq \rho^\star M \log N / N$ then for all $t > 0$, for all $N \geq N_0$, one has with probability at least $1 - Ce^{-t}$, for any permutation $\tau \in \mathfrak{S}_X$,*

$$\|\widehat{g} - g^\star\|_2^2 \leq 6 \inf_M \left\{ \|g^\star - g^{\mathbf{Q}^\star, f_M^\star}\|_2^2 + \mathrm{pen}(N, M) \right\} + C\frac{x}{N} + C\left\{ 2\|\mathbf{Q}^\star - \mathbb{P}_\tau \widehat{\mathbf{Q}}_N \mathbb{P}_\tau^\top\|_F^2 + \|\pi^\star - \mathbb{P}_\tau \widehat{\pi}\|_2^2 \right\}.$$

Here, $\mathbb{P}_\tau$ is the permutation matrix associated to τ (label switching).

Key step(s) of the proof: We use concentration inequalities for dependent variables [Pau15]. The proof strategy is based on a peeling argument. Here the model is not a vector space and we have to work finely using bracket entropy computations to catch (up to a log factor) the true complexity MX (instead of M^3 as for the spectral method) of the model $\mathscr{S}_{\mathbf{Q},M}$. ■

Now we have to derive an oracle inequality on the emission densities from the aforementioned result. In **[DC8]** the strategy was to lower bound $\|g^{\mathbf{Q},f}-g^{\mathbf{Q},f^\star}\|_2$ by $\|f-f^\star\|_2$ for $\mathbf{Q}$ and f in neighborhoods of $\mathbf{Q}^\star$ and $f^\star$. To achieve this result, we had to develop an *ad hoc* argument based on positive definiteness of some quadratic form. It results in the following assumption.

Assumption (G)

The transition matrix $\mathbf{Q}^\star$ and the emission densities $f^\star$ satisfy Assumption **(G)** if and only if $H(\mathbf{Q}^\star,(\langle f_k^\star, f_\ell^\star\rangle)_{k,\ell\in[X]}) \neq 0$ for some universal non constant polynomial H.

Assumption **(G)** is generically satisfied. In the case $X=2$, one can prove **[DC8]** that Assumption **(G)** is always satisfied.

Theorem 19 (**[DC8]**). *Assume **(F)**, **(G)** and **(H)**. Then there exist positive constants N_0, $\rho^\star$, C such that, if* $\mathrm{pen}(N,M) \geq \rho^\star M \log N/N$ *then for all $t>0$, for all $N \geq N_0$, one has with high probability, for any permutation $\tau_N \in \mathfrak{S}_X$, there exists a permutation $\tau' \in \mathfrak{S}_X$ such that*

$$\sum_{x=1}^{X}\|f^\star_{\tau'(x)}-\widehat{f}_{\tau_N(x)}\|_2^2 \leq C\Big[\inf_M\Big\{\sum_{x=1}^{X}\|f^\star_x-f^\star_{M,x}\|_2^2+\mathrm{pen}(N,M)\Big\}+\|\mathbf{Q}^\star-\mathbb{P}_{\tau_N}\widehat{\mathbf{Q}}\mathbb{P}_{\tau_N}^\top\|_F^2+\|\pi^\star-\mathbb{P}_{\tau_N}\widehat{\pi}\|_2^2+\frac{x}{N}\Big].$$

Here, $\mathbb{P}_{\tau_N}$ is the permutation matrix associated to τ_N (label switching).

An important consequence of this theorem is that the right choice of penalty leads to a rate minimax adaptive estimator up to a $\log N$ term, see Corollary 3 below. For this purpose, one has to choose an estimator $\widehat{\mathbf{Q}}$ of $\mathbf{Q}^\star$ which is consistent (up to label switching) with controlled rate. One possible choice is a spectral estimator. Indeed, the spectral estimator with, for each N, the dimension M_N chosen such that[2] $\eta(\Phi_{M_N}) = O((\log N)^{\frac{1}{4}})$, gives the following result.

Corollary 3. *With this choice of $\widehat{\mathbf{Q}}$, under the assumptions of Theorem 19, there exists a sequence of permutations $\tau_N \in \mathfrak{S}_X$ such that as N tends to infinity,*

$$\mathbb{E}\Big[\sum_{x=1}^{X}\|f^\star_x-\widehat{f}_{\tau_N(x)}\|_2^2\Big] = O\Big(\inf_{M'}\Big\{\sum_{x=1}^{X}\|f^\star_x-f^\star_{M',x}\|_2^2+\mathrm{pen}(N,M')\Big\}+\frac{\log N}{N}\Big).$$

Thus, choosing $\mathrm{pen}(N,M) = \rho M \log N/N$ for a large $\rho > 0$ leads to the minimax asymptotic rate of convergence up to a power of $\log N$. Indeed, standard results in approximation theory[3] show that one can upper bound the approximation error $\|f_k^\star - f_{M,k}^\star\|_2$ by $\mathscr{O}(M^{-\frac{s}{d}})$ where $s>0$ denotes a minimal regularity parameter. Then the optimal trade-off is obtained for $M^{\frac{1}{d}} \sim (N/\log N)^{\frac{1}{2s+d}}$, which leads to the quasi-optimal rate $(N/\log N)^{-\frac{s}{2s+d}}$ for the nonparametric estimation when the minimal smoothness of the emission densities is s. Notice that the algorithm automatically selects the best M leading to this rate.

3.1.3 Spectral method

The spectral method based on "*observable operators*" may have first occurred in brain imaging papers, see for instance [Jae00, JZK+07]. The key idea is to form an "*observable*" matrix from observed quantities in such manner that its spectrum estimates the emission laws,

[2] See Section 3.1.3 for a definition of η.
[3] See [DL93] for instance.

see Lemma 20. More precisely, we aim at estimating the coefficients of the emission densities $f_x^\star$ onto the orthonormal basis Φ_M given by, for all $(m,x)\in[M]\times[X]$,

$$\mathbf{A}_M(m,x) := \mathbb{E}(\varphi_m(Y_1)|X_1=x) = \langle f_x^\star, \varphi_m\rangle,$$

This idea was successfully used in [AHK12, HKZ12] in the parametric frame and, following the same guidelines, in **[DC8]** and **[DC14]** in the nonparametric frame. The observed quantities involved in the spectral methods are empirical joint laws, namely for all $(a,b,c)\in[M]^3$, for all $(m,x)\in[M]\times[X]$,

$$\mathbf{P}_{123}(a,b,c) := \mathbb{E}(\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)),$$
$$\mathbf{P}_{13}(a,c) := \mathbb{E}(\varphi_a(Y_1)\varphi_c(Y_3)).$$

Note that $\mathbf{P}_{13}$ and $\mathbf{P}_{123}$ can be estimated by their empirical counterparts, and $\mathbf{A}_M$ is the target matrix to estimate. The next lemma is the key result introducing the "*observable*" operators $\mathbf{B}(m)$ for all $m\in[M]$.

Lemma 20. *Let U be any $M\times X$ matrix such that $\mathbf{P}_{13}U$ has full rank X. Then $U^\top\mathbf{P}_{13}U$ is invertible and there exists an invertible matrix R such that for all $m\in[M]$,*

$$\mathbf{B}(m) := (U^\top\mathbf{P}_{13}U)^{-1}U^\top\mathbf{P}_{123}U = R\,\mathrm{Diag}[\mathbf{A}_M(m,.)]\,R^{-1}$$

where $\mathrm{Diag}[\mathbf{A}_M(m,.)]$ *is the diagonal matrix with diagonal entries* $(\mathbf{A}_M(m,x))_{x\in[X]}$.

The matrix U can be taken as the right singular matrix of $\mathbf{P}_{13}$ though other matrices are also acceptable. The proof of this lemma is rather elementary when writing the joint law matrices $\mathbf{P}_{13}$ and $\mathbf{P}_{123}$ is terms of $\mathbf{Q}^\star$, π and $\mathbf{A}_M$, see **[DC14]**. However, the result is essential, as it links the joint law matrices $\mathbf{P}_{13}$ and $\mathbf{P}_{123}$ (that can be efficiently estimated) to the target emission densities $\mathbf{A}_M$. An important feature of the observable operators $(\mathbf{B}(m))_{m\in[M]}$ is that they are jointly diagonalizable in the same[4] basis R.

Some "tricks" can be used to improve the efficiency of the spectral algorithm, presented in Algorithm 2. We will not comment on each of them though we will point to the use of the matrices

$$\mathbf{C}(x) := \sum_{m=1}^{M} U(m,x)\mathbf{B}(m)$$

indexed by $x\in[X]$. On can prove that

$$\mathbf{C}(x) = R\mathrm{Diag}[U^\top\mathbf{A}_M(m,.)]R^{-1},$$

as well. Hence these matrices satisfy the same benefits as the matrices $\mathbf{B}(m)$. In particular, they are also jointly diagonalizable onto the basis R. This basis can be estimated by diagonializing the empirical version of $\mathbf{C}(1)$ and re-used with the estimates of $\mathbf{C}(x)$ for $x\geq 2$. The trick is that we have only X matrices $\mathbf{C}(x)$ while there are M matrices $\mathbf{B}(m)$ where M, the dimension of the approximation space, can be large. So the algorithm is more robust when using the same estimation of the diagonalization basis but on fewer matrices, see Step 6 in Algorithm 2.

Now we introduce a last notation to state the rate of convergence of the spectral estimator. Using standard results [Pau15] on Bernstein concentration for depend observations, we consider the following variance bound,

$$\eta^2(\Phi_M) := \sup_{y,y'\in\mathscr{Y}^3}\sum_{a,b,c=1}^{M}\left(\varphi_a(y_1)\varphi_b(y_2)\varphi_c(y_3)-\varphi_a(y_1')\varphi_b(y_2')\varphi_c(y_3')\right)^2. \tag{3.4}$$

[4] Indeed the diagonalization matrix R does not depend on $m\in[M]$, see Lemma 20.

Algorithm 2: Spectral estimation of the transition matrix and the emission laws

Data: An observed chain $(Y_1, \dots, Y_{N+2})$ and a number of hidden states X.
Result: Spectral estimators $\widehat{\pi}$, $\widehat{\mathbf{Q}}$ and $(\widehat{f}_{M,x})_{x\in[X]}$.

1. For all a, b, c in $[M]$, consider the following empirical estimators:
$$\widehat{\mathbf{P}}_1(a) := \frac{1}{N}\sum_{s=1}^{N}\varphi_a(Y_s),\ \widehat{\mathbf{P}}_{123}(a,b,c) := \frac{1}{N}\sum_{s=1}^{N}\varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2}),$$
$$\widehat{\mathbf{P}}_{12}(a,b) := \frac{1}{N}\sum_{s=1}^{N}\varphi_a(Y_s)\varphi_b(Y_{s+1}) \text{ and } \widehat{\mathbf{P}}_{13}(a,c) := \frac{1}{N}\sum_{s=1}^{N}\varphi_a(Y_s)\varphi_c(Y_{s+2}).$$
2. Let $\widehat{U}$ be the $M \times X$ matrix of orthonormal right singular vectors of the matrix $\widehat{\mathbf{P}}_{13}$ corresponding to its top X singular values.
3. For all $m \in [M]$, set $\widehat{\mathbf{B}}(m) := (\widehat{U}^\top \widehat{\mathbf{P}}_{13}\widehat{U})^{-1}\widehat{U}^\top \widehat{\mathbf{P}}_{123}(.,m,.)\widehat{U}$.
4. Draw uniformly Θ a $(X \times X)$ unitary matrix and set $\widehat{\mathbf{C}}(x) := \sum_{m=1}^{M}(\widehat{U}\Theta)(m,x)\widehat{\mathbf{B}}(m)$, for all $x \in [X]$.
5. Compute $\widehat{R}$ a $(X \times X)$ unit Euclidean norm columns matrix that diagonalizes the matrix $\widehat{\mathbf{C}}(1)$ and set $\widehat{R}^{-1}\widehat{\mathbf{C}}(1)\widehat{R} = \mathrm{Diag}(\widehat{\Lambda}(1,1), \dots, \widehat{\Lambda}(1,K))$.
6. For all $x, x' \in [X]$, set $\widehat{\Lambda}(x,x') := (\widehat{R}^{-1}\widehat{\mathbf{C}}(x)\widehat{R})(x',x')$ and $\widehat{\mathbf{A}}_M := \widehat{U}\Theta\widehat{\Lambda}$.
7. Set the estimator $(\widehat{f}_{M,x})_{x\in[X]}$ defined by $\widehat{f}_{M,x} := \sum_{m=1}^{M}\widehat{\mathbf{A}}_M(m,x)\varphi_m$, for all $x \in [X]$.
8. Set $\widetilde{\pi} := \left(\widehat{U}^\top\widehat{\mathbf{A}}_M\right)^{-1}\widehat{U}^\top\widehat{\mathbf{P}}_1$.
9. Consider the estimator
$$\widehat{\mathbf{Q}} := \Pi_{\mathrm{TM}}\left(\left(\widehat{U}^\top\widehat{\mathbf{A}}_M \mathrm{Diag}\widetilde{\pi}\right)^{-1}\widehat{U}^\top\widehat{\mathbf{P}}_{12}\widehat{U}\left(\widehat{\mathbf{A}}_M^\top\widehat{U}\right)^{-1}\right),$$
where Π_{TM} denotes the projection (with respect to the scalar product given by the Frobenius norm) onto the convex set of transition matrices, and define $\widehat{\pi}$ as the stationary distribution of $\widehat{\mathbf{Q}}$.

Note that in classical examples (splines, Fourier, wavelets) one has

$$\eta(\Phi_M) \leq C M^{\frac{3}{2}},$$

for some numerical $C > 0$. To control the performances of our estimators, we use the quadratic loss. It can be expressed as a variance term and a bias term as follows,

$$\forall x \in [X],\ \forall M \geq 0, \quad \|f_x^\star - \widehat{f}_x\|_2^2 = \|\widehat{f}_x - f_{M,x}^\star\|_2^2 + \|f_x^\star - f_{M,x}^\star\|_2^2,$$

where $f_{M,x}^\star$ is the projection of $f_x^\star$ onto $\mathfrak{P}_M$ and $\widehat{f}_x$ is any estimator such that $\widehat{f}_x \in \mathfrak{P}_M$. Note that the bias term $\|f_x^\star - f_{M,x}^\star\|_2$ does not depend on the estimator $\widehat{f}_x$, it comes from the approximation properties of the basis Φ_M and decreases with the complexity M. The variance term $\min_{\tau\in\mathfrak{S}_X}\max_{x\in[X]}\|\widehat{f}_x - f_{M,\tau(x)}^\star\|_2^2$ accounts for the performances of the estimator $\widehat{f}_k$ and increases with M. As usual in nonparametric statistics, a good choice of M has to balance these two terms. The next result gives a bound on the variance term of the spectral estimator depending on both the sample size N and the complexity M.

Theorem 21 ([DC8] and [DC14]). *Assume **(H)**. There exists a numerical constant $C>0$ that may depend on $\mathbf{Q}^\star$ and $\mathfrak{F}^\star$ such that the following holds. For any $t>C$, for any $\delta\in(0,1)$, for any $M\geq C$, there exists a permutation $\tau_M\in\mathfrak{S}_X$ such that, for any $N\geq C\eta(\Phi_M)^2 t(-\log\delta)/\delta$,*

$$\|f^\star_{M,x}-\widehat{f}_{M,\tau_M(x)}\|_2\leq C\frac{\sqrt{-\log\delta}}{\delta}\frac{\eta(\Phi_M)}{\sqrt{N}}\sqrt{t},$$
$$\|\pi^\star-\mathbb{P}_{\tau_M}\widehat{\pi}\|_2\leq C\frac{\sqrt{-\log\delta}}{\delta}\frac{\eta(\Phi_M)}{\sqrt{N}}\sqrt{t},$$
$$\|\mathbf{Q}^\star-\mathbb{P}_{\tau_M}\widehat{\mathbf{Q}}\mathbb{P}^\top_{\tau_M}\|\leq C\frac{\sqrt{-\log\delta}}{\delta}\frac{\eta(\Phi_M)}{\sqrt{N}}\sqrt{t},$$

with probability greater than $1-2\delta-4e^{-t}$. Furthermore, choosing a sequence $(M_N)_N$ of integers tending to infinity and such that $\eta(\Phi_{M_N})=o(\sqrt{N/\log N})$, there exists a sequence of permutations $\tau_N\in\mathfrak{S}_X$ such that

$$\mathbb{E}\big[\max_{x\in[X]}\|f^\star_{M_N,x}-\widehat{f}_{\tau_N(x)}\|_2^2\big]\vee\mathbb{E}\big[\|\mathbf{Q}^\star-\mathbb{P}_{\tau_N}\widehat{\mathbf{Q}}\mathbb{P}^\top_{\tau_N}\|^2\big]\vee\mathbb{E}\big[\|\pi^\star-\mathbb{P}_{\tau_N}\widehat{\pi}\|_2^2\big]=O\Big(\eta(\Phi_{M_N})^2\frac{\log N}{N}\Big)=o(1).$$

Here, the expectations are with respect to the observations and to the random unitary matrix Θ drawn at Step 4 of Algorithm 2.

Key step(s) of the proof: The proof makes an intensive use of perturbation matrix theory for controlling eigenvalues and eigenvectors under a small perturbation. The dependence in M has been carefully tracked to achieve the present result. ■

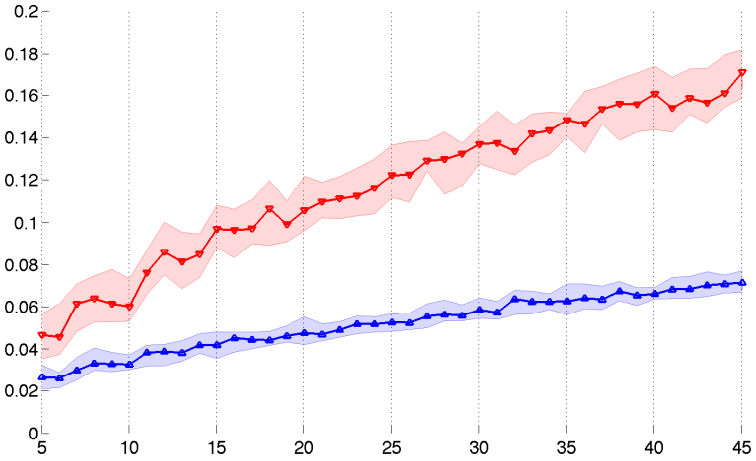

Figure 3.1: Variance of the spectral (red) and empirical least-square (blue) estimators.

Given any function ω such that $\omega(N)/\log N$ tends to infinity (as slowly as desired) as N tends to infinity, note that if one chooses M_N such that $\eta(\Phi_{M_N})^2=\omega(N)/\log N$ then there exists a sequence of permutations $\tau_N\in\mathfrak{S}_X$ such that

$$\mathbb{E}\big[\|\mathbf{Q}^\star-\mathbb{P}_{\tau_N}\widehat{\mathbf{Q}}\mathbb{P}^\top_{\tau_N}\|^2\big]\leq O\Big(\frac{\omega(N)}{N}\Big)$$

which is almost the parametric rate of estimation. Furthermore, Theorem 21 states that the variance term is of order M^3/N (recall that in standard cases $\eta(\Phi_M)=O(M^{\frac{3}{2}})$). The balance between bias and variance is then achieved for $N\sim M^{3+2s/d}$ that leads to the rate $N^{-\frac{s}{2s+3d}}$ for the nonparametric spectral estimation though the minimax rate is $N^{-\frac{s}{2s+d}}$. The rate of the spectral estimator is the standard rate of approximation of a density in dimension $3d$ with smoothness s. This may reflect that the spectral estimator is based on the estimation of

the joint density of three consecutive observations (Y_1, Y_2, Y_3). To get a variance term of the minimax order M/N we have deployed the penalized least-squares method that captures the intrinsic complexity of the statistical model for the distribution of (Y_1, Y_2, Y_3) that is[5] XM.

As depicted in Figure 3.1, we have compared, for each M, the variance terms obtained by the spectral method and the empirical least-squares method over 40 iterations on chains of length $N = 5e4$. For each curve, we have plotted a shaded box plot representing the first and third quartiles. We have considered $X = 2$ hidden states whose emission variables are distributed with respect to beta laws of parameters $(2,5)$ and $(4,2)$. This numerical experiment consolidates the idea that the least squares method significantly improves upon the spectral method. Indeed, even for small values of M, one may see in Figure 3.1 that the variance term is divided by at least a constant factor.

3.1.4 Estimation algorithm

Recall that the experimenter knows nothing about the underlying hidden Markov model but the number of hidden states X. Our procedure is based on the computation of the empirical least-squares estimators $\widehat{g}_M$ defined as minimizers of the empirical contrast γ_N on the space $\mathscr{S}(\widehat{\mathbf{Q}}, M)$ where $\widehat{\mathbf{Q}}$ is an estimator of the transition matrix (for instance the spectral estimation of the transition matrix). Since the function γ_N is non convex, we use a second order approach estimating a positive definite matrix (using a covariance matrix) within an iterative procedure called CMAES for Covariance Matrix Adaptation Evolution Strategy, see [Han06]. Using this latter algorithm, we search for the minimum of γ_N taking as a starting point the spectral estimation of the emission laws. Then, we estimate the size of the model thanks to

$$\widehat{M}(\rho) \in \arg\min_{M=1,\dots,M_{\max}} \left\{ \gamma_N(\widehat{g}_M) + \rho \frac{M \log N}{N} \right\},$$

where the penalty term ρ has to be tuned and the maximum size of the model $M_{\max}$ can be set by the experimenter in a data-driven procedure.

Indeed, we shall apply the slope heuristic to adjust the penalty term and to choose $M_{\max}$. As presented in [BMM12], the minimum contrast function $M \mapsto \gamma_N(\widehat{g}_M)$ should have a linear behavior for large values of M. The experimenter has to consider $M_{\max}$ large enough in order to observe this linear stabilization. The slope of the linear interpolation is then $(\widehat{\rho}/2)\log N/N$ (recall that the sample size N is fixed here) where $\widehat{\rho}$ is the slope heuristic choice on how ρ should be tuned. Another procedure (theoretically equivalent) consists in plotting the function $\rho \mapsto \widehat{M}(\rho)$ which is a non-increasing piecewise constant function. The estimated $\widehat{\rho}$ is such that the largest drop (called "*dimension jump*") of this function occurs at point $\widehat{\rho}/2$, see **[DC8]** for some numerical examples. To summarize, our procedure reads as follows.

1. For all $M \leq M_{\max}$, compute the spectral estimations $(\widehat{\mathbf{Q}}, \widehat{\pi})$ of the transition matrix and its stationary distribution and the spectral estimation $\widetilde{f}$ of the emission laws. This is straightforward using the procedure described by Algorithm 2.

2. For all $M \leq M_{\max}$, compute a minimum $\widehat{g}_M$ of the empirical contrast function γ_N using "Covariance Matrix Adaptation Evolution Strategy", see [Han06]. Use the estimation $\widetilde{f}$ of the spectral method as a starting point of CMAES.

3. Tune the penalty term using the slope heuristic procedure and select $\widehat{M}$.

4. Return the emission laws of the solution of point (2) for $M = \widehat{M}$.

Note that the size M of the projection space for the spectral estimator has been set as the one chosen by the slope heuristic for the empirical least squares estimators.

In Figure 3.2, we consider the regular histogram basis and the trigonometric basis for estimating emission laws given by beta laws of parameters $(2,5)$ and $(4,2)$ from a single chain

[5] Indeed, we have X emission distributions with M degrees of freedom.

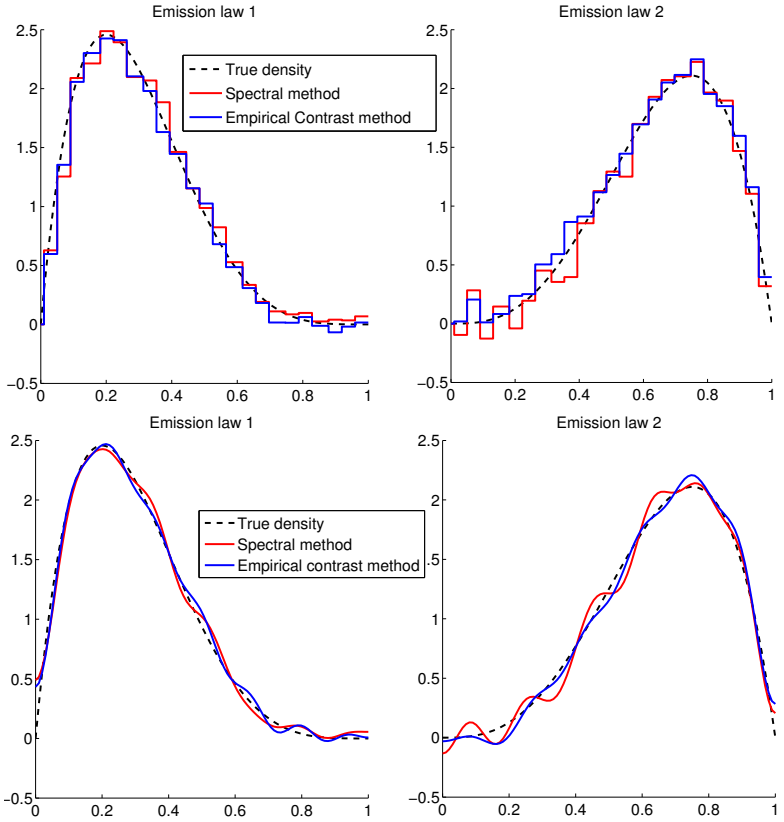

Figure 3.2: Spectral and penalized least-squares estimators of the emission densities.

observation of length $N = 5e4$. On the top panels, we have used the histogram basis ($\widehat{M} = 23$). On the bottom panels, we have considered the trigonometric basis ($\widehat{M} = 21$).

3.1.5 Nonparametric filtering and smoothing

In many applications of finite state space HMMs (*e.g.,* digital communication or speech recognition), it is of utmost importance to infer the sequence of hidden states. The filtering and smoothing tasks focus on the underlying hidden states chain while estimating the conditional probabilities

$$\phi^{\star}_{k}(x, y_{1:k}) := \mathbb{P}\{X_k = x | Y_{1:k} = y_{1:k}\} \qquad \text{(Filtering)}$$
$$\phi^{\star}_{k|N}(x, y_{1:N}) := \mathbb{P}\{X_k = x | Y_{1:N} = y_{1:N}\} \qquad \text{(Smoothing)}$$

where $x \in [X]$ and $Y_{1:k} := (Y_1, \ldots, Y_k)$ for $k \geq 1$. In **[DC14]**, we study how the parameter estimation error propagates to the error made on the estimation of filtering and smoothing distributions.

Although replacing parameters by their estimators to compute posterior distributions and infer the hidden states is usual in applications, theoretical results to support this practice are very few regarding the accuracy of the estimated posterior distributions. We are only aware of [EDKM07] whose results are restricted to the filtering distribution in a parametric setting. When the parameters of the HMM are known, the forward-backward algorithm can be extended to general state space HMMs or when the cardinality of $[X]$ is too large using computational methods such as Sequential Monte Carlo methods (SMC). In this context, the Forward Filtering Backward Smoothing and Forward Filtering Backward Simulation algorithms have been intensively studied, with the objective of quantifying the error made when the filtering and marginal smoothing distributions are replaced by their Monte Carlo approximations. These algorithms and some extensions have recently been analyzed theoretically, see

for instance [DMDS10, DGM+11]. SMC methods may also be used in algorithms when the parameters of the HMM are unknown to perform maximum likelihood parameter estimation, see [KDS+15] for on-line and off-line Expectation Maximization and gradient ascent based algorithms. Part of our analysis of the filtering and smoothing distributions is based on the same approach and requires sharp forgetting properties of HMMs.

For all $y_{1:N} \in \mathscr{Y}^N$, the filtering distributions $\phi^\star_k(\cdot, y_{1:k})$ and marginal smoothing distributions $\phi^\star_{k|N}(\cdot, y_{1:N})$ may be computed explicitly for all $k \in [N]$ using the forward-backward algorithm of [BPSW70]. In the forward pass, the filtering distributions $\phi^\star_k$ are updated recursively using the identities, for all $x \in [X]$,

$$\phi^\star_1(x, y_1) := \frac{\pi^\star(x) f^\star_x(y_1)}{\sum_{x' \in [X]} \pi^\star(x') f^\star_{x'}(y_1)} \text{ and } \phi^\star_k(x, y_{1:k}) := \frac{\sum_{x' \in [X]} \mathbf{Q}^\star(x', x) f^\star_x(y_k) \phi^\star_{k-1}(x', y_{1:k-1})}{\sum_{x', x'' \in [X]} \mathbf{Q}^\star(x', x'') f^\star_{x''}(y_k) \phi^\star_{k-1}(x', y_{1:k-1})}.$$

In the backward pass, the marginal smoothing distributions may be updated recursively using, for all $x \in [X]$,

$$\phi^\star_{N|N}(x, y_{1:N}) := \phi^\star_N(x, y_{1:N}) \quad \text{and} \quad \phi^\star_{k|N}(x, y_{1:N}) := \sum_{x'=1}^{X} B^\star_{\phi^\star_k(\cdot, y_{1:k})}(x', x) \phi^\star_{k+1|N}(x', y_{1:n}),$$

where, for all $u, v \in [X]$ and all $k \in [N]$,

$$B^\star_{\phi^\star_k(\cdot, y_{1:k})}(u, v) := \frac{\mathbf{Q}^\star(v, u) \phi^\star_k(v, y_{1:k})}{\sum_{z \in [X]} \mathbf{Q}^\star(z, u) \phi^\star_k(z, y_{1:k})}.$$

Since the parameters $\pi^\star$, $\mathbf{Q}^\star$ and $f^\star$ are unknown, the aforementioned recursive equations may be applied to some estimators $\widehat{\pi}$, $\widehat{\mathbf{Q}}$ and $\widehat{f}$ to obtain approximations of the filtering and smoothing distributions.

A standard proof's technique on "*forgetting properties*" assumes that the transitions are lower bounded, namely

$$\delta^\star := \min_{1 \le i, j \le X} \mathbf{Q}^\star(i, j) > 0. \qquad \textbf{(Assumption (I))}$$

Similarly, we denote $\widehat{\delta} := \min_{1 \le i, j \le X} \widehat{\mathbf{Q}}(i, j)$. We assume that we are given a set of $N = p + n$ observations from the hidden Markov model driven by $\pi^\star$, $\mathbf{Q}^\star$ and $f^\star$. The first p observations are used to produce the estimators $\widehat{\pi}$, $\widehat{\mathbf{Q}}$ and $\widehat{f}$ while filtering and smoothing are performed with the last n observations. In other words the estimators $\widehat{\pi}$, $\widehat{\mathbf{Q}}$ and $\widehat{f}$ are measurable functions of $Y_{1:p}$ and the objective is to estimate $\phi^\star_k(\cdot, Y_{p+1:p+k})$ and $\phi^\star_{k|n}(\cdot, Y_{p+k:p+n})$. In the following, we denote by $\|\cdot\|_1$ the total variation norm.

Theorem 22 ([DC14]). *Assume **(H)** and **(I)**. Then for all $n \ge 1$, for any permutation $\tau_p \in \mathfrak{S}_X$,*

$$\sup_{1 \le k \le n} \mathbb{E}\Big[\|\phi^\star_k(\cdot, Y_{p+1:p+k}) - \widehat{\phi}^{\tau_p}_k(\cdot, Y_{p+1:p+k})\|_1\Big] \le \frac{C_\star}{(\delta^\star)^2} \Big\{\mathbb{E}[\|\pi^\star - \mathbb{P}_{\tau_p} \widehat{\pi}_p\|_2] + \mathbb{E}[\|\mathbf{Q}^\star - \mathbb{P}_{\tau_p} \widehat{\mathbf{Q}}_p \mathbb{P}^\top_{\tau_p}\|_F] + \sum_{x=1}^{X} \mathbb{E}[\|f^\star_x - \widehat{f}_{\tau_p(x)}\|_1]\Big\}$$

and, for the smoothing part,

$$\sup_{1 \le k \le n} \mathbb{E}\Big[\|\phi^\star_{k|n}(\cdot, Y_{p+1:p+n}) - \widehat{\phi}^{\tau_p}_{k|n}(\cdot, Y_{p+1:p+n})\|_1\Big] \le \frac{C_\star}{(\delta^\star)^2} \Big\{\mathbb{E}[\|\pi^\star - \mathbb{P}_{\tau_p} \widehat{\pi}_p\|_2] + \mathbb{E}\Big[\|\mathbf{Q}^\star - \mathbb{P}_{\tau_p} \widehat{\mathbf{Q}}_p \mathbb{P}^\top_{\tau_p}\|_F / \widehat{\delta}\Big] + \sum_{x=1}^{X} \mathbb{E}\Big[\|f^\star_x - \widehat{f}_{\tau_p(x)}\|_1 / \widehat{\delta}\Big]\Big\}.$$

Here, $\rho_\star := 1 - \delta^\star/(1 - \delta^\star)$, $C_\star := 4(1 - \delta^\star)/\delta^\star$, and $\widehat{\phi}^{\tau_p}_k$ and $\widehat{\phi}^{\tau_p}_{k|n}$ are the estimations of $\phi^\star_k$ and $\phi^\star_{k|n}$ based on $\mathbb{P}_{\tau_p} \widehat{\mathbf{Q}} \mathbb{P}^\top_{\tau_p}$, $\mathbb{P}_{\tau_p} \widehat{\pi}$ and $\widehat{f}_{\tau_p(x)}$, for all $x \in [X]$.

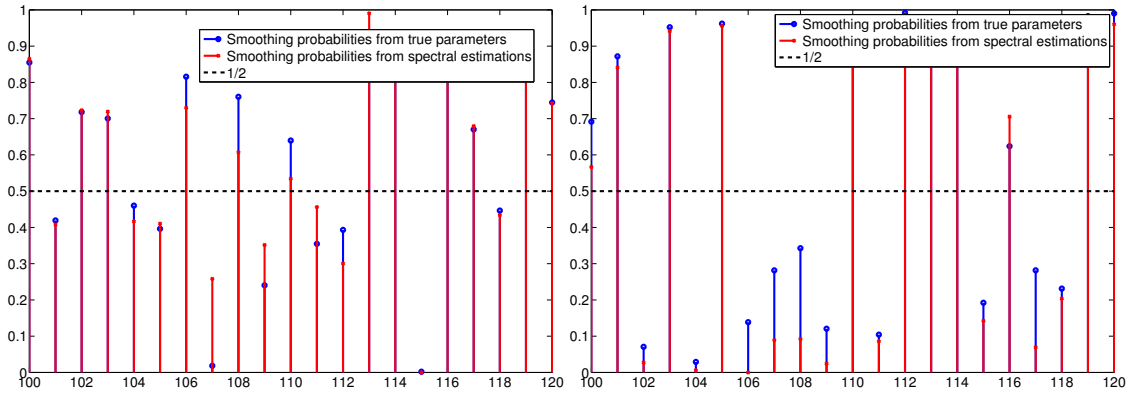

Figure 3.3: Marginal smoothing probabilities with the spectral method.

Key step(s) of the proof: Using forgetting properties of the hidden chain, we are able to obtain an upper bound of the filtering errors and of the marginal smoothing errors by terms involving only the estimation errors of $\pi^\star$, $\mathbf{Q}^\star$ and $f^\star$. Then, we derive a fine control on how the estimation errors made on the parameters of the HMM propagate upon the filtering and smoothing distributions. ■

We end by setting the result that follows when using the spectral estimator. Let $(M_p)_{p\geq 1}$ be an increasing sequence of integers. Recall the definition of the orthonormal approximation basis Φ_{M_p} seen in Section 3.1.2 and the concentration variance bound $\eta(\Phi_{M_p})$ viewed in (3.4).

Corollary 4 ([DC14])**.** *Assume **(H)** and **(I)**. Choosing a sequence $(M_p)_p$ of integers tending to infinity and such that $\eta(\Phi_{M_p}) = o(\sqrt{p/\log p})$, there exists a sequence of permutations $\tau_p \in \mathfrak{S}_X$ such that*

$$\mathbb{E}\Big[\sup_{k\geq 1} \|\phi_k^\star(\cdot, Y_{p+1:p+k}) - \widehat{\phi}_k^{\tau_p}(\cdot, Y_{p+1:p+k})\|_1\Big] = O\Big(\eta(\Phi_{M_p})\sqrt{\frac{\log p}{p}} + \sum_{x=1}^{X} \|f_x^\star - f_{M_p,x}^\star\|_2\Big)$$

and, for the smoothing part,

$$\mathbb{E}\Big[\sup_{1\leq k\leq n} \|\phi_{k|n}^\star(\cdot, Y_{p+1:p+n}) - \widehat{\phi}_{k|n}^{\tau_p}(\cdot, Y_{p+1:p+n})\|_1\Big] = O\Big(\eta(\Phi_{M_p})\sqrt{\frac{\log p}{p}} + \sum_{x=1}^{X} \|f_x^\star - f_{M_p,x}^\star\|_2\Big).$$

Here, the expectations are with respect to the observations and to the random unitary matrix Θ drawn at Step 4 of Algorithm 2.

Recall that in standard cases (splines, Fourier, wavelets) one has $\eta(\Phi_{M_p}) = O(M_p^{\frac{3}{2}})$. As in Section 3.1.3, the balance between bias and variance is then achieved for $p \sim M^{3+2s/d}$ that leads to the rate $p^{-\frac{s}{2s+3d}}$ for the nonparametric spectral estimation of the posterior probabilities. Indeed, recall that the rate of the spectral estimator is the standard rate of approximation of a density in dimension $3d$ with smoothness s, see Section 3.1.3.

We have run several numerical experiments to assess the efficiency of our method. We consider $X = 2$ emission laws of beta distributions with parameters $(2,5)$ and $(4,3)$. We observe a sequence of $N = 6e4$ observations $(Y_i)_{i=1}^N$ from which we use the $p = 5e4$ first observations to estimate the parameters $\pi^\star$, $\mathbf{Q}^\star$ and $f^\star$ using the spectral method. As projection basis, we have considered the histogram basis (left panel) or the trigonometric basis (right panel) in Figure 3.3. From the spectral method estimates, we compute an estimation of the marginal smoothing probabilities using the forward-backward algorithm.

3.2 Reconstructing undirected graphs from eigen spaces

In **[DC15]**, we consider a new set of problems where one aims at recovering an undirected weighted graph of size N from an estimation of the eigen spaces of its adjacency matrix W and incomplete information on its set of edges[6]. This situation depicts any model where one knows in advance a linear operator K that commutes with W. Several examples and the general model is given in Section 3.2.1. In particular, we assume that we have access to an estimation $\widehat{\mathsf{K}}$ of K build from an n-sample and we consider the empirical contrast given by the "*commutator*", namely $\mathsf{A} \mapsto \|\widehat{\mathsf{K}}\mathsf{A} - \mathsf{A}\widehat{\mathsf{K}}\|_2$ where $\|\cdot\|_2$ denotes the Frobenius norm. Using backward-type procedures based on this empirical contrast, Section 3.2.2 derives estimators of the graph structure, *i.e.,* its set of edges $S^\star$, also called "*support*". This study reveals typical behaviors of the empirical contrast when the estimated support S contains or not the true support $S^\star$. A thresholding heuristic is developed in Section 3.2.3.

3.2.1 Model and identifiability

Consider a symmetric matrix $\mathsf{W} \in \mathbb{R}^{N\times N}$ giving the adjacency matrix of an undirected weighted graph on N vertices. We focus on the eigen spaces of W examining models where we have no information on the spectrum of the graph. Depicting this situation, we assume that we consider a matrix $\mathsf{K} \in \mathbb{R}^{N\times N}$ such that $\mathsf{KW} = \mathsf{WK}$ or, in more realistic scenarios, we may observe a perturbed version $\widehat{\mathsf{K}}$ of K. The key point is then to use extra information given by the location of some zero entries of W. Hence, we assume that one knows in advance a set $F \subset [N]^2$ of "*forbidden*" entries such that

$$\forall (i,j) \in F, \quad \mathsf{W}_{ij} = 0 \tag{$\mathbf{H}_\mathrm{F}$}$$

Equivalently, the set F is disjoint from the set of edges of the target graph. For $S \subseteq [N]^2$, denote by $\mathscr{E}(S)$ the set of symmetric matrices A whose support is included in S, namely $\mathrm{supp}(\mathsf{A}) \subseteq S$. Given the set F of forbidden entries defined via ($\mathbf{H}_\mathrm{F}$), the matrix of interest W is sought in the set $\mathscr{E}(\overline{F})$ where $\overline{F}$ denotes the complement of F. In some cases, most matrices $\mathsf{W} \in \mathscr{E}(\overline{F})$ are uniquely determined by their eigen spaces. More precisely, for each of those $\mathsf{W} \in \mathscr{E}(\overline{F})$, there is no matrix $\mathsf{A} \in \mathscr{E}(\overline{F})$ non colinear with W that commutes with W. This property is encapsulated by the notion of *F-identifiability* as follows.

Definition (F-identifiability)**.** *We say that a matrix W is F-identifiable if, and only if, the only solutions A with $\mathrm{supp}(\mathsf{A}) \subseteq \overline{F}$ to $\mathsf{AW} = \mathsf{WA}$ are of the form $\mathsf{A} = \lambda \mathsf{W}$ for some $\lambda \in \mathbb{R}$.*

Interestingly, we have the following proposition.

Proposition 23 (Lemma 2.1 in **[DC4]**)**.** *Let $S \subseteq \overline{F}$, the set of F-identifiable matrices in $\mathscr{E}(S)$ is either empty or a dense open subset of $\mathscr{E}(S)$.*

Key step(s) of the proof: The proof uses the fact that non F-identifiable matrices in $\mathscr{E}(S)$ can be expressed as the zeros of a particular analytic function using determinants. ■

This proposition shows that the F-identifiability of a matrix W is essentially a condition on its support S. By abuse of notation, we say that a support $S \subseteq \overline{F}$ is *F-identifiable* if almost every matrix in $\mathscr{E}(S)$ are F-identifiable. Characterizing the F-identifiability appears to be a challenging issue since it can be viewed as understanding the eigen structure of a graph through its support.

Denote by $F = F_{\mathrm{diag}} := \{(i,i),\ 1 \le i \le N\}$ the set of forbidden entries representing that there are no self-loops in W. The F_{diag}-identifiability, or *diagonal identifiability*, can be reasonably assumed in many practical situations since it entails that W lives on a simple graph, with no self-loops. In **[DC15]**, we introduce necessary and sufficient conditions on the target support $\mathrm{supp}(\mathsf{W})$ for diagonal identifiability. Defining the *kite* graph[7] ∇_N of size N as the graph (V,E)

[6] For example, one knows that the target graph has no self-loops.

[7] See Figure 3.4 for instance.

with vertices $V = [N]$ and edges $E = \{(k,k+1),\ 1 \leq k \leq N-1\} \cup \{(N-2,N)\}$, one simple sufficient condition on diagonal identifiability reads as follows.

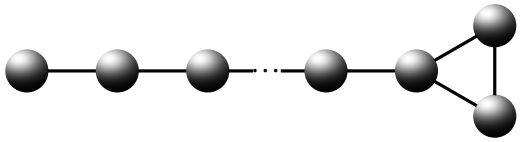

Figure 3.4: The kite graph ∇_N.

Proposition 24 ([DC15]). *If the kite graph ∇_N of size N is a subgraph of the graph of size N and edges S then S is diagonally identifiable.*

Denote $G(N,p)$ the Erdős-Rényi model on graphs of size N where the edges are drawn independently with respect to the Bernoulli law of parameter p. One can prove [DC15] that $\log N/N$ is a sharp threshold for diagonal identifiability in the Erdős-Rényi model, it can be stated as follows.

Theorem 25 ([DC15]). *Diagonal identifiability in the Erdős-Rényi model occurs with a sharp phase transition with threshold function $\log N/N$: for any $\varepsilon > 0$, it holds*

- *If $p_N \geq (1+\varepsilon)\log N/N$ and $G_N \sim G(N,p_N)$ then the probability that $\mathrm{supp}(G_N)$ is diagonally identifiable tends to 1 as N goes to infinity.*
- *If $p_N \leq (1-\varepsilon)\log N/N$ and $G_N \sim G(N,p_N)$ then the probability that $\mathrm{supp}(G_N)$ is diagonally identifiable tends to 0 as N goes to infinity.*

Key step(s) of the proof: This result relies on a companion result provided in [DC15] that gives a necessary and a sufficient condition for identifiability. As shown in Proposition 24, it suffices to prove that a kite exists. This event is described by the first point and it can be elementarily proved using a decomposition on two independent Erdős-Rényi graphs. The converse is based on observing that it is sufficient to find two isolated vertices to prove non-identifiability. One knows (see in [Bol98, Theorem 3.1] for instance) that the event "there is at least two isolated points" has sharp threshold function $\log N/N$. It proves the second point. ■

In practice, one may expect that any target graph of size N with no self-loops and degree bounded from below by $\log N$ is diagonally identifiable. In this case, it might be recovered from its eigen spaces. Conversely, small degree graphs (*i.e.,* graphs with some vertices of degree much smaller than $\log N$) may not be identifiable. In this case, there is no hope to reconstruct it from its eigen spaces since there exists another small degree undirected weighted graph with the same eigen spaces.

Some models that can be treated by our approach cover Markov chains observed at i.i.d. random time gaps with unknown law [DC4], Vectorial AutoRegressive process observed at random time gaps as well, Graphical models, and Seasonal VAR structure, see [DC15].

3.2.2 Estimating the support

Several approaches can be unleashed to estimate the target support $S^\star$ though standard convex relaxation techniques penalizing W may fail since W is scale invariant. A first approach is given by penalizing the ℓ_0-norm of the support and choosing an appropriate criterion. This approach satisfies pleasant theoretical properties but it meets with the curse of dimensionality [DC15], especially since the size of the support increases quadratically with the dimension. In practice, a backward methodology provides a computationally feasible alternative to the support reconstruction problem. Starting from the maximal acceptable support $\overline{F}$, the idea of the backward procedure is to remove the least significant entries one at a time and stop when every entry is significant.

Using the corresponding small case letter to denote the vectorization of a matrix, *e.g.*, we write $a = \mathrm{vec}(\mathsf{A}) = (\mathsf{A}_{11},\dots,\mathsf{A}_{N1},\dots,\mathsf{A}_{1N},\dots,\mathsf{A}_{NN})^\top$, significancy can be leveraged using the Frobenius norm of the commutator operator $a \mapsto \Delta(\mathsf{K})a = \mathrm{vec}(\mathsf{KA}-\mathsf{AK})$, where we denote $\Delta(\mathsf{K}) = \mathrm{I}\otimes\mathsf{K}-\mathsf{K}\otimes\mathrm{I} \ \in \mathbb{R}^{N^2\times N^2}$ with $\otimes$ the Kronecker product. Indeed, searching for the target W in the commutant of K amounts to searching for $w = \mathrm{vec}(\mathsf{W})$ in $\ker(\Delta(\mathsf{K}))$, the kernel of $\Delta(\mathsf{K})$. Note that the functions $\mathsf{A} \mapsto \|\widehat{\mathsf{K}}\mathsf{A}-\mathsf{A}\widehat{\mathsf{K}}\|_2^2$ and $a \mapsto \|\Delta(\widehat{\mathsf{K}})a\|_2^2$ can be used indistinctly as cost functions. Minimizing this criterion over model spaces of decreasing size, we consider sequences of least-squares estimates in the sequel. This empirical criterion was first used in **[DC4]** and **[DC15]** to reflect that W is expected to nearly commute with $\widehat{\mathsf{K}}$, provided that $\widehat{\mathsf{K}}$ is sufficiently close to its true value K. We assume the following hypotheses ($\mathbf{H}_\Sigma$), ($\mathbf{H}_1$) and ($\mathbf{H}_{\mathrm{Id}}$).

◦ Deriving the asymptotic law of least-squares estimators, we may assume that the estimate $\widehat{\mathsf{K}}$ is built from a sample X of size n growing to infinity and asymptotically Gaussian with an asymptotic covariance matrix either known or that can be estimated. One can write

$$\sqrt{n}(\widehat{k}-k) \xrightarrow[n\to\infty]{d} \mathcal{N}_{N^2}(0,\Sigma), \tag{$\mathbf{H}_\Sigma$}$$

where Σ is a $N^2\times N^2$ covariance matrix. This condition is verified for instance in the framework considered in **[DC4]**, **[DC15]** and [BPR16]. Note that asymptotic normality is a standard ground base when investigating any least-square procedure.

◦ In order to exclude the trivial solution $a = 0$, the target W is assumed normalized so that

$$\mathbf{1}^\top w = 1, \tag{$\mathbf{H}_1$}$$

where $\mathbf{1}$ has all its entries equal to one. Because the available information on W is of spectral nature and as such, is scale-invariant, a normalization of some kind is crucial for the identifiability. Here, the condition $\mathbf{1}^\top w = 1$ achieves two goals: preventing the null matrix form being a solution and making the problem identifiable[8] .

◦ For S a support included in $\overline{F}$, we aim at a solution in the affine space

$$\mathscr{A}_S := \{a = \mathrm{vec}(\mathsf{A}) : \mathrm{supp}(\mathsf{A}) \subseteq S,\ \mathsf{A} = \mathsf{A}^\top,\ \mathbf{1}^\top a = 1\}.$$

with linear difference space given by $\mathscr{L}_S := \{a = \mathrm{vec}(\mathsf{A}) : \mathrm{supp}(\mathsf{A}) \subseteq S,\ \mathsf{A} = \mathsf{A}^\top,\ \mathbf{1}^\top a = 0\}$. By abuse of notation, $\mathscr{A}_S$ may refer both to the space of matrices or their vectorizations. To find the target support $S^\star$, one must exploit the fact that the vector w lies in the intersection of $\ker(\Delta(\mathsf{K}))$ and $\mathscr{A}_{\overline{F}}$. Actually, w can then be recovered if the intersection is reduced to the singleton $\{w\}$. In this case, the matrix W and its support $S^\star$ are F-identifiable. Hence, we assume that

$$\ker(\Delta(\mathsf{K}))\cap\mathscr{L}_{\overline{F}} = \{0\}, \tag{$\mathbf{H}_{\mathrm{Id}}$}$$

which is implied by F-identifiability, see Definition 3.2.1.

Consider for each support $S \subseteq \overline{F}$ a full-ranked $N^2\times\dim(\mathscr{A}_S)$ matrix Φ_S whose column vectors form a basis of $\mathscr{L}_S$. Since W is F-identifiable and $S \subseteq \overline{F}$, the operator $\Delta(\mathsf{K})\Phi_S$ is one-to-one. Denoting by $\mathsf{M}^\dagger$ the Moore-Penrose pseudoinverse of a matrix M, we consider the estimator $\widehat{w}_S = \mathrm{vec}(\widehat{\mathsf{W}}_S) = \arg\min_{a\in\mathscr{A}_S}\|\Delta(\widehat{\mathsf{K}})a\|_2^2$ where $\Omega_S = (\Phi_S^\top\Delta(\mathsf{K}))^\dagger\Delta(\mathsf{W})\Sigma\Delta(\mathsf{W})(\Delta(\mathsf{K})\Phi_S)^\dagger$. One can prove **[DC15]** that

$$\sqrt{n}(\widehat{w}_s - w) \xrightarrow[n\to\infty]{d} \mathcal{N}_{N^2}\big(0,\Phi_S\Omega_S\Phi_S^\top\big), \tag{3.5}$$

[8]The main drawback of this normalization concerns the situation where the entries of W sum up to zero, in which case the normalization is impossible. If the context suggests that the solution may be such that $\mathbf{1}^\top w = 0$, a different affine normalization $\mathbf{v}^\top w = 1$ (with any fixed vector $\mathbf{v}$) must be used, without major changes in the methodology. In practice, one may consider the vector $\mathbf{v}$ as random (for instance with isotropic law), so that ($\mathbf{H}_1$) is almost surely fulfilled for any fixed target w.

the asymptotic distribution of $\widehat{w}_S$. The limit covariance matrix is unknown, but plugging the estimates $\widehat{\mathsf{W}}_S$, $\widehat{\mathsf{K}}$ and $\widehat{\Sigma}$ yields a consistent estimator $\Phi_S\widehat{\Omega}_S\Phi_S^\top$. So, denotting $\widehat{\sigma}^2_{S,ij}$ the diagonal entry of $\Phi_S\widehat{\Omega}_S\Phi_S^\top$ associated to $\widehat{\mathsf{W}}_{S,ij}$, the edge (i,j) is judged significant if the statistic

$$\tau_{ij}(S) := \sqrt{n}\frac{\widehat{\mathsf{W}}_{S,ij}}{\widehat{\sigma}_{S,ij}} \tag{3.6}$$

exceeds in absolute value some quantile of the standard Gaussian distribution. The backward support selection procedure is then implemented by the recursive algorithm, see Algorithm 3.

Algorithm 3: Thresholded backward algorithm for support selection

Data: A set of forbidden entries F, a matrix $\widehat{\mathsf{K}}_n$, a threshold $t>0$.
Result: An estimated support S.

1: Start with the maximal acceptable support $S_1 = \overline{F}$,
2: At each Step m, compute the statistics $\tau_{ij}(S_m)$ for all $(i,j)\in S_m$,
3: If the minimal value $|\tau_{ij}(S_m)|$ is smaller than t, set $S_{m+1} = S_m \setminus \{(i,j),(j,i)\}$,
4: Stop when all entries are judged significant, *i.e.,* when $|\tau_{ij}(S_m)| > t$.

Based on (3.5), the quantile $q_{1-\frac{\alpha}{2}}$ of the standard Gaussian distribution appears as a reasonable choice for the threshold, as it boils down to performing an asymptotic significativity test of level α. However, due to the slow convergence to the limit distribution and the tendency to overestimate the variance for small sample sizes, the numerical study of Section 3.2.3 shows that a better choice for the threshold depends on the overall behavior of the commutator $\Delta(\widehat{\mathsf{K}})\widehat{w}_{S_m}$ computed over the nested sequence of active supports. The details are discussed in Section 3.2.3 where the backward procedure is presented.

3.2.3 The boosted backward algorithm for support selection

An other procedure is based on the empirical contrast (criterion)

$$\forall S\subseteq\overline{F},\quad \mathrm{Crit}(S,\widehat{\mathsf{K}}) := \min_{\mathsf{A}\in\mathscr{E}(S)\setminus\{0\}}\frac{\|\mathsf{A}\widehat{\mathsf{K}}-\widehat{\mathsf{K}}\mathsf{A}\|_2^2}{\|\mathsf{A}\|_2^2}, \tag{3.7}$$

where $\mathscr{E}(S)$ is the set of symmetric matrices A whose support is included in S as defined in Section 3.2.1. Using this criterion, the ℓ_0-approach of **[DC15]** leads to a consistent estimation of the target support $S^\star$. Unfortunately, computing this estimator is NP-hard and one can not use it in practice. On the other hand, Section 3.2.2 offers a significance test and introduces the thresholded backward algorithm for support selection in Algorithm 3. The significance of an edge (i,j) of an active support S is evaluated by the test statistic $\tau_{ij}(S)$ defined in (3.6). Note that this test statistic is asymptotically Gaussian. However, in practice, the stopping condition based on the Gaussian quantiles and given in Section 3.2.2 happens to overestimate the support. Indeed, it does not take into account the fact that the same sample is used to remove edges from the active set. Furthermore, the use of Gaussian quantiles for the statistic τ_{ij} may be hazardous.

Using the best of the two aforementioned approaches, we can introduce a new backward type procedure. We begin by removing the least significant edge at each step, building a sequence of nested active supports $S_1\supset\cdots\supset S_\ell$, that we refer to as a "*trajectory*". Along this trajectory, we compute the empirical contrast defined by

$$\forall S\subseteq\overline{F},\quad S\mapsto\mathrm{Crit}(\widehat{\mathsf{W}}_S,\widehat{\mathsf{K}}) := \frac{\|\widehat{\mathsf{W}}_S\widehat{\mathsf{K}}-\widehat{\mathsf{K}}\widehat{\mathsf{W}}_S\|}{\|\widehat{\mathsf{W}}_S\|}. \tag{3.8}$$

When the true support $S^\star$ lies in the trajectory, one expects to observe a "gap" in the sequence $j\mapsto\mathrm{Crit}(\widehat{\mathsf{W}}_{S_j},\widehat{\mathsf{K}})$ when S_j goes from $S^\star$ to a smaller support.

Indeed:

- For $S^\star \subseteq S$, the target W is consistently estimated by $\widehat{\mathsf{W}}_S$ so that $\mathrm{Crit}(\widehat{\mathsf{W}}_S, \widehat{\mathsf{K}})$ tends to zero at rate $\sqrt{n}$,
- For $S \subsetneq S^\star$, the lower bound $\|\mathsf{A}\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\mathsf{A}\| \geq \|\mathsf{A}\mathsf{K} - \mathsf{K}\mathsf{A}\| - 2\|\widehat{\mathsf{K}} - \mathsf{K}\|\|\mathsf{A}\|$ yields

$$\mathrm{Crit}(\widehat{\mathsf{W}}_S, \widehat{\mathsf{K}}) = \frac{\|\widehat{\mathsf{W}}_S\widehat{\mathsf{K}} - \widehat{\mathsf{K}}\widehat{\mathsf{W}}_S\|}{\|\widehat{\mathsf{W}}_S\|} \geq c(S) - 2\|\widehat{\mathsf{K}} - \mathsf{K}\| \tag{3.9}$$

  with $c(S) := \min_{A \in \mathscr{A}_S} \|\mathsf{A}\mathsf{K} - \mathsf{K}\mathsf{A}\|/\|\mathsf{A}\|$ a positive constant. In particular, one has

$$\min_{S \subsetneq S^\star} c(S) \geq \min_{\substack{S \neq S^\star \\ |S| \leq |S^\star|}} c(S) =: c_0(S^\star) > 0$$

  where the right hand size term is positive by identifiability.

In some way, $c_0(S^\star)$ measures the amplitude of the signal: one expects to be able to recover the target W when the estimation error $\|\widehat{\mathsf{K}} - \mathsf{K}\|$ reaches at least the same order as $c_0(S^\star)$. The true support $S^\star$ then corresponds to a transitional gap in the contrast curve that can be captured by a suitably chosen threshold $t > 0$. Since $\widehat{\mathsf{K}}$ converges toward K in probability, any threshold $0 < t < c_0(S^\star)$ will work with probability one asymptotically.

An obstacle to the detection of the commutation gap is the increasing behavior of the commutator over the nested trajectory $S_1 \supset \cdots \supset S_\ell$. This phenomenon, indirectly caused by the dependence between the trajectory and $\widehat{\mathsf{K}}$, can be annihilated when considering the empirical contrast over a trajectory built from a training sample. In fact, the monotonicity can even be "reversed" before reaching the true support if the $\widehat{\mathsf{W}}_{S_j}$ are estimated independently from $\widehat{\mathsf{K}}$. Thus, the sequence $j \mapsto \mathrm{Crit}(\widetilde{\mathsf{W}}_{S_j}, \widehat{\mathsf{K}}) = \|\Delta(\widehat{\mathsf{K}})\widetilde{w}_{S_j}\|/\|\widetilde{w}_{S_j}\|$ is expected to achieve its minimum for the best estimator $\widetilde{w}_{S_j}$ in the trajectory, that is for $S_j = S^\star$. Furthermore, beyond the true support (for small active supports), $\widetilde{w}_{S_j}$ is not a consistent estimator of w so that the criterion no longer approaches zero, resulting in the so-called commutation gap.

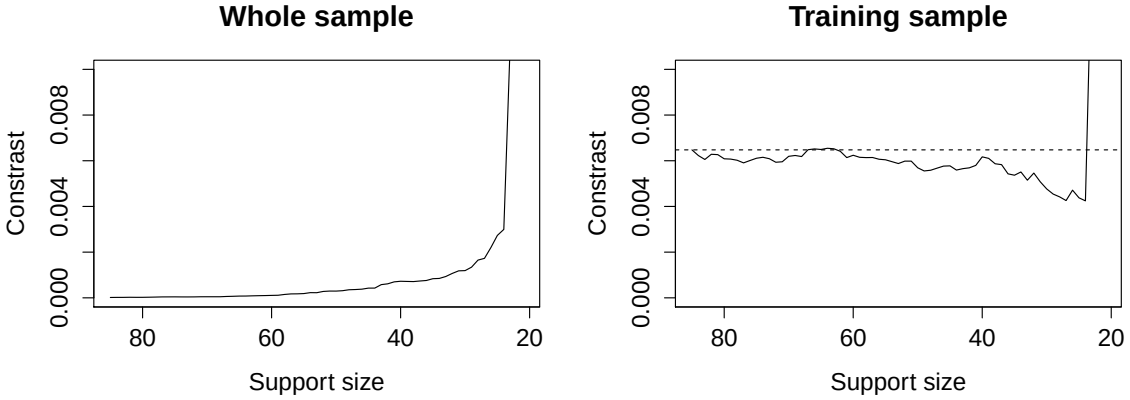

Figure 3.5: The contrast sequence $j \mapsto \mathrm{Crit}(\widetilde{\mathsf{W}}_{S_j}, \widehat{\mathsf{K}})$ is computed along a trajectory. The nested support sequence and estimators $\widetilde{\mathsf{W}}_{S_j}$ are obtained from the backward algorithm implemented on the whole sample (left) and on a training sample of half size (right). In both cases, $\widehat{\mathsf{K}}$ is constructed from the whole sample. Using a training sample manages to reverse the monotonicity in the first part of the sequence, thus making the commutation gap easier to locate. The initial value of the sequence $t = \mathrm{Crit}(\widetilde{\mathsf{W}}_{S_1}, \widehat{\mathsf{K}})$ then provides a tractable adaptive choice for the threshold.

The "reversed" monotonicity provides an easy way to calibrate the threshold in the backward algorithm. Indeed, since $S_j \mapsto \Delta(\widehat{\mathsf{K}})\widetilde{w}_{S_j}$ is expected to decrease when approaching the true support (coming from larger active supports along a trajectory), the estimated support can be heuristically chosen has the last time the criterion is below an adaptive threshold, see Figure 3.5. In particular, $\mathrm{Crit}(\widetilde{\mathsf{W}}_{S_1}, \widehat{\mathsf{K}})$ can be used as an adaptive threshold for the backward algorithm when the estimator $\widehat{\mathsf{K}}$ and the trajectory $S_1 \supset \cdots \supset S_\ell$ are obtained from independent samples.

Of course, to afford splitting the sample to build the $\widetilde{\mathsf{W}}_{S_j}$ independent from $\widehat{\mathsf{K}}$ may be unrealistic. Nevertheless, the numerical study suggests that the independence is well mimicked when $\widehat{\mathsf{K}}$ is built from the whole dataset but the backward algorithm sequence $\widetilde{\mathsf{W}}_{S_1}, ..., \widetilde{\mathsf{W}}_{S_\ell}$ is obtained from a learning sub-sample, as illustrated in Figure 3.5. Empirically, the optimal size of training samples could be calibrated in function of the number of observations using the robustness of the outputs of the algorithm. In this paper, we always draw training samples by taking each observation with probability $1/2$, with no consideration regarding the size of the whole sample.

From a computational cost point of view, evaluating Criterion (3.8) at an active support S recasts in computing the smallest eigenvalue of the operator $\Delta(\widehat{\mathsf{K}})$ on the space $\mathscr{E}(S)$ of size $|S|^2$, see Section 3.2.2 for definitions. Given a trajectory, selecting an active support falls into the frame of spectral methods on a space of size at most $|\overline{F}|^2 = \mathscr{O}(N^4)$ where N denotes the number of vertices of the graph. If N is greater than 50, cheaper criteria should be used to reduce the computational cost but we did not pursue in this direction.

3.3 Prospects

1. Latent space models have recently flourished and rapidly attracted a lot of attention in the statistical and machine learning communities. This competitive area has numerous offsprings. While the general "graphon" model has been recently studied, little statistical analysis has been achieved compared to the broad variety of models, and important statistical questions are still open. This year, I started a project **[DC20]** with some colleagues on nonparametric estimation of graphon in some particular random graph models. The model we are studying has not been treated yet from the nonparametric perspective. This work raises important and interesting questions of nonparametric estimation of graphons when specifying the latent space, *e.g.,* spheres in our case. This teamwork will undoubtedly open new prospects of research for future work.

2. While working on **[DC15]** exciting combinatorial questions attracted our attention. One of them can be summarized as follows. Can we recover a graph from the knowledge of the number of paths of length k starting and ending at the same vertex v for all $k \in \mathbb{N}$ and all vertex v? If not, what can be characterized by those numbers?

Notations

$\#S$	Size of the set S
$(\mathbb{K}, d)$	Compact metric set
$[d]$	Set $[d] = \{1, \dots, d\}$
$\mathbf{E}$	Real-valued continuous functions over $\mathbb{K}$ endowed with the supremum norm
δ_t	Dirac mass at point $t \in \mathbb{K}$
$\mathbb{R}_{++}$	Set of positive real numbers
$\mathbf{C}_n$	Cone of moments
$\mathbf{M}$	Markov system, see Section 2.1
$\mathscr{E}(S)$	Set of symmetric matrices A whose support is included in S
$\mathscr{N}_d(\mu, \Theta)$	d-dimensional Gaussian law with mean μ and variance Θ
$\mathscr{P}_1$	Set of coefficients of bounded generalized polynomials, see (2.3) for a definition
$\mathfrak{S}_X$	Set of permutations of $[X]$
$\mathbb{S}_p^+$	Space of symmetric nonnegative definite matrices of size p
$\mathbf{E}^\star$	Banach space of real Borel measures endowed with the total variation norm
Σ_s	Space of s-sparse vectors
n	Number of measurements
S^c	Complement of S
X	Design matrix
LMI	Linear Matrix Inequality, see (2.5)

Bibliography

[AH16] Grigory Alexandrovich and Hajo Holzmann, *Nonparametric identification of hidden markov models*, Biometrika (2016).

[AHK12] Animashree Anandkumar, Daniel Hsu, and Sham M Kakade, *A method of moments for mixture models and hidden markov models*, 25th Annual Conference on Learning Theory, vol. 23, 2012, pp. 33.1–33.34.

[ALMT14] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp, *Living on the edge: Phase transitions in convex programs with random data*, Information and Inference: a Journal of the IMA **3** (2014), no. 3, 224–294.

[AMR09] Elizabeth S Allman, Catherine Matias, and John A Rhodes, *Identifiability of parameters in latent structure models with many observed variables*, The Annals of Statistics (2009), 3099–3132.

[Ant10] Anestis Antoniadis, *Comments on: ℓ_1-penalization for mixture regression models*, Test **19** (2010), no. 2, 257–258.

[AW09] Jean-Marc Azaïs and Mario Wschebor, *Level sets and extrema of random processes and fields*, John Wiley & Sons Inc., 2009.

[BC11] Heinz H Bauschke and Patrick L Combettes, *Convex analysis and monotone operator theory in hilbert spaces*, Springer Science & Business Media, 2011.

[BCT11] Jeffrey D Blanchard, Coralia Cartis, and Jared Tanner, *Compressed sensing: How sharp is the restricted isometry property?*, SIAM review **53** (2011), no. 1, 105–125.

[BCW11] Alexandre Belloni, Victor Chernozhukov, and Lie Wang, *Square-root lasso: pivotal recovery of sparse signals via conic programming*, Biometrika **98** (2011), no. 4, 791–806.

[BDF14] Tamir Bendory, Shai Dekel, and Arie Feuer, *Exact recovery of non-uniform splines from the projection onto spaces of algebraic polynomials*, Journal of Approximation Theory **182** (2014), 7–17.

[BDF15a] ———, *Exact recovery of dirac ensembles from the projection onto spaces of spherical harmonics*, Constructive Approximation **42** (2015), no. 2, 183–207.

[BDF15b] ———, *Super-resolution on the sphere using convex optimization*, Signal Processing, IEEE Transactions on **63** (2015), no. 9, 2253–2262.

[BDF16] ———, *Robust recovery of stream of pulses using convex optimization*, Journal of Mathematical Analysis and Applications (2016).

[Beu38] Arne Beurling, *Sur les intégrales de fourier absolument convergentes et leur application à une transformation fonctionnelle*, Ninth Scandinavian Mathematical Congress, 1938, pp. 345–366.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, OUP Oxford, 2013.

[BLPR11] Karine Bertin, Erwan Le Pennec, and Vincent Rivoirard, *Adaptive dantzig density estimation*, Ann. Inst. H. Poincaré Probab. Statist. **47** (2011), no. 1, 43–74.

[BMM12] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel, *Slope heuristics: overview and implementation*, Statistics and Computing **22** (2012), no. 2, 455–470.

[Bol98] Béla Bollobás, *Random graphs*, Springer, 1998.

[BP13] Kristian Bredies and Hanna Katriina Pikkarainen, *Inverse problems in spaces of measures*, ESAIM: Control, Optimisation and Calculus of Variations **19** (2013), no. 01, 190–218.

[BPR16] Flavia Barsotti, Anne Philippe, and Paul Rochet, *Hypothesis testing for markovian models with random time observations*, Journal of Statistical Planning and Inference (2016), 87–98.

[BPSW70] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss, *A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains*, Annals of Mathematical Statistics **41** (1970), no. 1, 164–171.

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov, *Simultaneous analysis of lasso and Dantzig selector*, Ann. Statist. **37** (2009), no. 4, 1705–1732.

[BS10] Zhidong Bai and Jack W. Silverstein, *Spectral analysis of large dimensional random matrices*, second ed., Springer Series in Statistics, Springer, New York, 2010.

[BvdG11] Peter Bühlmann and Sara A. van de Geer, *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media, 2011.

[CDD09] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore, *Compressed sensing and best k-term approximation*, J. Amer. Math. Soc. **22** (2009), no. 1, 211–231.

[CDS98] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput. **20** (1998), no. 1, 33–61.

[CFG13] Emmanuel J. Candès and Carlos Fernandez-Granda, *Super-resolution from noisy data*, Journal of Fourier Analysis and Applications **19** (2013), no. 6, 1229–1254.

[CFG14] ———, *Towards a mathematical theory of super-resolution*, Communications on Pure and Applied Mathematics **67** (2014), no. 6, 906–956.

[CGH+96] Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth, *On the Lambert W function*, Advances in Computational mathematics **5** (1996), no. 1, 329–359.

[CGLP12] Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor, *Interactions between compressed sensing, random matrices, and high dimensional geometry*, vol. 37, Société Mathématique de France, 2012.

[CP09] Emmanuel J. Candès and Yaniv Plan, *Near-ideal model selection by ℓ_1 minimization minimization*, The Annals of Statistics **37** (2009), no. 5A, 2145–2177.

[CRT06] Emmanuel J. Candès, Justin Romberg, and Terence Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory **52** (2006), no. 2, 489–509.

[CT05] Emmanuel J. Candès and Terence Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory **51** (2005), no. 12, 4203–4215.

[CT06] ———, *Near-optimal signal recovery from random projections: universal encoding strategies?*, IEEE Trans. Inform. Theory **52** (2006), no. 12, 5406–5425.

[CT07] ———, *The Dantzig selector: statistical estimation when p is much larger than n*, Ann. Statist. **35** (2007), no. 6, 2313–2351.

[DDP15] Quentin Denoyelle, Vincent Duval, and Gabriel Peyré, *Asymptotic of sparse support recovery for positive measures*, Journal of Physics: Conference Series, vol. 657, IOP Publishing, 2015.

[DET06] David L. Donoho, Michael Elad, and Vladimir N. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, Information Theory, IEEE Transactions on **52** (2006), no. 1, 6–18.

[DG96] Paul Doukhan and Fabrice Gamboa, *Superresolution rates in prokhorov metric*, Canadian Journal of Mathematics **48** (1996), no. 2, 316–329.

[DGM+11] Randal Douc, Aurélien Garivier, Eric Moulines, Jimmy Olsson, et al., *Sequential monte carlo smoothing for general state space hidden markov models*, The Annals of Applied Probability **21** (2011), no. 6, 2109–2145.

[DHL15] Arnak S. Dalalyan, Mohamed Hebiri, and Johannes Lederer, *On the prediction performance of the lasso*, Bernoulli (2015).

[DL93] Ronald A DeVore and George G Lorentz, *Constructive approximation*, vol. 303, Springer Science & Business Media, 1993.

[DMDS10] Pierre Del Moral, Arnaud Doucet, and Sumeetpal S Singh, *A backward particle interpretation of feynman-kac formulae*, ESAIM: Mathematical Modelling and Numerical Analysis **44** (2010), no. 05, 947–975.

[Don92] David L. Donoho, *Superresolution via sparsity constraints*, SIAM Journal on Mathematical Analysis **23** (1992), no. 5, 1309–1331.

[Don05] ———, *Neighborly polytopes and sparse solutions of underdetermined linear equations*, Tech. report, Stanford University, 2005.

[Don06a] ———, *Compressed sensing*, IEEE Trans. Inform. Theory **52** (2006), no. 4, 1289–1306.

[Don06b] ———, *High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension*, Discrete & Computational Geometry **35** (2006), no. 4, 617–652.

[DP15a] Vincent Duval and Gabriel Peyré, *Exact support recovery for sparse spikes deconvolution*, Foundations of Computational Mathematics **15** (2015), no. 5, 1315–1355.

[DP15b] ———, *The non degenerate source condition: Support robustness for discrete and continuous sparse deconvolution*, Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on, Dec 2015, pp. 49–52.

[DS97] Holger Dette and William J Studden, *The theory of canonical moments with applications in statistics, probability, and analysis*, vol. 338, John Wiley & Sons, 1997.

[DS01] Kenneth R Davidson and Stanislaw J Szarek, *Local operator theory, random matrices and banach spaces*, Handbook of the geometry of Banach spaces **1** (2001), no. 317-366, 131.

[DT05] David L. Donoho and Jared Tanner, *Neighborliness of randomly projected simplices in high dimensions*, Proceedings of the National Academy of Sciences of the United States of America **102** (2005), no. 27, 9452–9457.

[DT09a] ———, *Counting faces of randomly projected polytopes when the projection radically lowers dimension*, Journal of the American Mathematical Society **22** (2009), no. 1, 1–53.

[DT09b] ———, *Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing*, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **367** (2009), no. 1906, 4273–4293.

[Dum07] Bogdan Dumitrescu, *Positive trigonometric polynomials and signal processing applications*, Springer, 2007.

[EDKM07] Eyal Even-Dar, Sham M Kakade, and Yishay Mansour, *The value of observation for monitoring dynamic systems.*, IJCAI, 2007, pp. 2474–2479.

[EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani, *Least angle regression*, Ann. Statist. **32** (2004), no. 2, 407–499, With discussion, and a rejoinder by the authors.

[FG13] Carlos Fernandez-Granda, *Support detection in super-resolution*, The 10th International Conference on Sampling Theory and Applications (SampTA 2013), 2013, pp. 145–148.

[FG16] ———, *Super-resolution of point sources via convex programming*, Information and Inference: a Journal of the IMA **5** (2016), no. 3, 251–303.

[FGS15] Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon, *Mind the duality gap: safer rules for the lasso*, International Conference on Machine Learning, 2015.

[FL09] Simon Foucart and Ming-Jun Lai, *Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for* $0 < q \leq 1$, Applied and Computational Harmonic Analysis **26** (2009), no. 3, 395–407.

[FR13] Simon Foucart and Holger Rauhut, *A mathematical introduction to compressive sensing*, Springer, 2013.

[FS10] Ohad N Feldheim and Sasha Sodin, *A universality result for the smallest eigenvalues of certain sample covariance matrices*, Geometric And Functional Analysis **20** (2010), no. 1, 88–123.

[Fuc04] Jean-Jacques Fuchs, *On sparse representations in arbitrary redundant bases*, Information Theory, IEEE Transactions on **50** (2004), no. 6, 1341–1344.

[Fuc05] ———, *Recovery of exact sparse representations in the presence of bounded noise*, Information Theory, IEEE Transactions on **51** (2005), no. 10, 3601–3608.

[GCR16] Élisabeth Gassiat, Alice Cleynen, and Stephane Robin, *Inference in finite state space non parametric hidden markov models and applications*, Statistics and Computing **26** (2016), no. 1-2, 61–71.

[GG96] Fabrice Gamboa and Élisabeth Gassiat, *Sets of superresolution and the maximum entropy method on the mean*, SIAM journal on mathematical analysis **27** (1996), no. 4, 1129–1152.

[Gir14] Christophe Giraud, *Introduction to high-dimensional statistics*, vol. 138, CRC Press, 2014.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan, *Unbalanced expanders and randomness extractors from parvaresh–vardy codes*, Journal of the ACM (JACM) **56** (2009), no. 4, 20.

[Han06] Nikolaus Hansen, *The cma evolution strategy: a comparing review*, Towards a new evolutionary computation, Springer, 2006, pp. 75–102.

[HKZ12] Daniel Hsu, Sham M Kakade, and Tong Zhang, *A spectral algorithm for learning hidden markov models*, Journal of Computer and System Sciences **78** (2012), no. 5, 1460–1480.

[HTFF05] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin, *The elements of statistical learning: data mining, inference and prediction*, The Mathematical Intelligencer **27** (2005), no. 2, 83–85.

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright, *Statistical learning with sparsity: the lasso and generalizations*, CRC Press, 2015.

[Jae00] Herbert Jaeger, *Observable operator models for discrete stochastic time series*, Neural Computation **12** (2000), no. 6, 1371–1398.

[JN11] Anatoli Juditsky and Arkadi Nemirovski, *Accuracy guarantees for l1-recovery*, Information Theory, IEEE Transactions on **57** (2011), no. 12, 7818–7839.

[JZK+07] Herbert Jaeger, Mingjie Zhao, Klaus Kretzschmar, Tobias Oberstein, Dan Popovici, and Andreas Kolling, *Learning observable operator models via the es algorithm*, ch. 14, pp. 417–464, MIT Press, 2007.

[Kaš77] Boris Sergeevich Kašhin, *Diameters of some finite-dimensional sets and classes of smooth functions*, Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya **41** (1977), no. 2, 334–351.

[KDS+15] Nikolas Kantas, Arnaud Doucet, Sumeetpal S Singh, Jan Maciejowski, and Nicolas Chopin, *On particle methods for parameter estimation in state-space models*, Statistical science **30** (2015), no. 3, 328–351.

[Kie74] Jack Kiefer, *General equivalence theory for optimum designs (approximate theory)*, The annals of Statistics (1974), 849–879.

[KN77] Mark G. Krein and Adolf A. Nudelman, *The markov moment problem and extremal problems*, Translations of Mathematical Monographs, vol. 50, American Mathematical Society, 1977.

[Kru77] Joseph B Kruskal, *Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics*, Linear algebra and its applications **18** (1977), no. 2, 95–138.

[Las09] Jean-Bernard Lasserre, *Moments, positive polynomials and their applications*, vol. 1, World Scientific, 2009.

[LM16] Guillaume Lecué and Shahar Mendelson, *Sparse recovery under weak moment assumptions*, Journal of the European Mathematical Society (2016).

[Lou08] Karim Lounici, *Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators*, Electronic Journal of Statistics **2** (2008), 90–102.

[LR+10] Michel Ledoux, Brian Rider, et al., *Small deviations for beta ensembles*, Electronic Journal of Probability **15** (2010), no. 41, 1319–1343.

[LSST13] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor, *Exact post-selection inference with the lasso*, arXiv preprint arXiv:1311.6238 (2013).

[LTTT14] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani, *A significance test for the lasso*, Annals of statistics **42** (2014), no. 2, 413.

[MT14] Michael B McCoy and Joel A Tropp, *Sharp recovery bounds for convex demixing, with applications*, Foundations of Computational Mathematics **14** (2014), no. 3, 503–567.

[Nie14] Jiawang Nie, *Optimality conditions and finite convergence of lasserre's hierarchy*, Mathematical programming **146** (2014), no. 1-2, 97–121.

[Owe07] Art B Owen, *A robust hybrid of lasso and ridge regression*, Contemporary Mathematics **443** (2007), 59–72.

[Pau15] Daniel Paulin, *Concentration inequalities for markov chains by marton couplings and spectral methods*, Electronic Journal of Probability **20** (2015), no. 79, 1–32.

[PV05] Farzad Parvares and Alexander Vardy, *Correcting errors beyond the guruswami-sudan radius in polynomial time*, Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on, IEEE, 2005, pp. 285–294.

[RWY11] Garvesh Raskutti, Martin J Wainwright, and Bin Yu, *Minimax rates of estimation for high-dimensional linear regression over-balls*, Information Theory, IEEE Transactions on **57** (2011), no. 10, 6976–6994.

[SBvdG10] Nicolas Städler, Peter Bühlmann, and Sara A. van de Geer, ℓ_1*-penalization for mixture regression models*, TEST **19** (2010), no. 2, 209–256.

[Sch06] Claus Scheiderer, *Sums of squares on real algebraic surfaces*, manuscripta mathematica **119** (2006), no. 4, 395–410.

[Sto10] Mihailo Stojnic, ℓ_1 *optimization and its various thresholds in compressed sensing*, Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, IEEE, 2010, pp. 3910–3913.

[Sto13] ———, *A rigorous geometry-probability equivalence in characterization of* ℓ_1*-optimization*, arXiv preprint arXiv:1303.7287 (2013).

[SZ12] Tingni Sun and Cun-Hui Zhang, *Scaled sparse linear regression*, Biometrika (2012).

[TBR15] Gongguo Tang, Badri N. Bhaskar, and Benjamin Recht, *Near minimax line spectral estimation*, Information Theory, IEEE Transactions on **61** (2015), no. 1, 499–512.

[TBSR13] Gongguo Tang, Badri N. Bhaskar, Parikshit Shah, and Benjamin Recht, *Compressed sensing off the grid*, Information Theory, IEEE Transactions on **59** (2013), no. 11, 7465–7490.

[Tib96] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **58** (1996), no. 1, 267–288.

[Tik15] Konstantin Tikhomirov, *The limit of the smallest singular value of random matrices with iid entries*, Advances in Mathematics **284** (2015), 1–20.

[TLT13] Jonathan Taylor, Joshua Loftus, and Ryan Tibshirani, *Tests in adaptive regression via the kac-rice formula*, arXiv preprint arXiv:1308.3020 (2013).

[TLTT14] Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani, *Exact post-selection inference for forward stepwise and least angle regression*, arXiv preprint arXiv:1401.3889 **7** (2014).

[vdG16] Sara A. van de Geer, *Estimation and Testing under Sparsity*, Saint Flour Lecture Notes, Springer, 2016.

[vdGB09] Sara A. van de Geer and Peter Bühlmann, *On the conditions used to prove oracle results for the Lasso*, Electron. J. Stat. **3** (2009), 1360–1392.

[Wai09] Martin J Wainwright, *Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso)*, IEEE transactions on information theory **55** (2009), no. 5, 2183–2202.

[Wel74] Lloyd R. Welch, *Lower bounds on the maximum cross correlation of signals (corresp.)*, Information Theory, IEEE Transactions on **20** (1974), no. 3, 397–399.

[XCM10] H. Xu, C. Caramanis, and S. Mannor, *Robust regression and lasso*, IEEE Trans. Inf. Theory **56** (2010), no. 7, 3561–3574.

[XH08] Weiyu Xu and Babak Hassibi, *Compressed sensing over the grassmann manifold: A unified analytical framework*, Communication, Control, and Computing, 2008 46th Annual Allerton Conference on, IEEE, 2008, pp. 562–567.

[XH11] ———, *Precise stability phase transitions for minimization: A unified geometric framework*, Information Theory, IEEE Transactions on **57** (2011), no. 10, 6894–6919.

[ZY06] Peng Zhao and Bin Yu, *On model selection consistency of lasso*, The Journal of Machine Learning Research **7** (2006), 2541–2563.