Contribution to the study of graphical models and high-dimensional statistics applied to the modeling of Triple-Negative Breast Cancer

Eunice Okome Obiang

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

université angers

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ D'ANGERS
en collaboration avec l'Institut de Cancérologie de l'Ouest
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : Mathématiques

Par
Eunice OKOME OBIANG

Contribution to the study of graphical models and high-dimensional statistics applied to the modeling of Triple-Negative Breast Cancer

Thèse présentée et soutenue à l'Université d'Angers, le 18 Octobre 2022
Unités de recherche : LAREMA (UA)

Rapporteuse et rapporteur avant soutenance :

Julien Chiquet	Chercheur, Université Paris-Saclay, AgroParisTech, INRAE
Pierre Pudlo	Professeur, Institut de Mathématiques de Marseille

Composition du Jury :

Président :	Gilles Stupfler	Professeur, Université d'Angers
Examinatrices :	Malgorzata Bogdan	Professeure, Uniwersytet Wroclawski
	Agnès Lagnoux	Maître de Conférences, Université Toulouse Jean Jaurès
Dir. de thèse :	Loïc Chaumont	Professeur, Université d'Angers
Encadrants :	Frédéric Proïa	Maître de Conférences, Université d'Angers
	Pascal Jézéquel	Praticien hospitalier

Résumé

Contribution à l'étude des modèles graphiques et statistique en grande dimension appliquée à la modélisation du cancer du sein triple négatif

Eunice Okome Obiang, Ph.D.
Université d'Angers en collaboration avec l'Institut de Cancérologie de l'Ouest, 2022

Cette thèse s'articule autour de deux axes. Le premier constitue une contribution à l'étude des modèles graphiques gaussiens partiels (PGGM) dans le cadre de l'apprentissage en grande dimension. Plus précisément, nous nous intéressons à la modélisation à sorties multiples, où nous souhaitons estimer d'une part la matrice Δ des liens directs entre les prédicteurs et les réponses, et d'autre part la matrice de précision conditionnelle des réponses Ω_y. Nous débutons avec une approche fréquentiste par maximum de vraisemblance pénalisée, où nous proposons un PGGM muni de deux formes de pénalisation : une pénalisation ℓ_1 induisant de la sparsité sur Δ et Ω_y, et une pénalisation structurante reflétant un a priori gaussien généralisé sur les liens directs. Nous montrons que, lorsqu'il est convenablement régularisé, ce modèle est agrémenté d'une garantie théorique prenant la forme d'une borne supérieure sur l'erreur d'estimation. Enfin, nous clôturons cette première réflexion par des études empiriques mettant en avant le caractère structurant de cette procédure d'estimation, et sa pertinence sur un jeu de données réelles. Nous poursuivons par l'étude de la contrepartie bayésienne, jusqu'alors inexplorée dans la littérature. En suivant une stratégie spike and slab, nous offrons plusieurs structures hiérarchiques imposant soit une configuration saturée, sparse, group-sparse ou encore sparse-group-sparse de la matrice Δ. Nous obtenons une garantie théorique pour les configurations sparse et group-sparse, et illustrons les résultats compétitifs de ces modèles sur une étude de simulation et un jeu de données réels, menés avec des échantillonneurs de Gibbs. Le deuxième axe de la thèse est, quant à lui, entièrement dévolu à la sélection de variables pronostiques en analyse de suvie multi-omique. Nous y proposons un algorithme de sélection de variables descendante offrant un consensus entre différentes méthodes de régularisation, notamment celles présentées dans le premier axe. L'efficacité de cette approche est enfin étudiée sur des données relatives au cancer du sein triple négatif, en prenant le soin de répondre aux contraintes identifiées par les oncologues. Tous nos codes sont rendus disponibles à la communauté.

Abstract

Contribution to the study of graphical models and high-dimensional statistics applied to the modeling of Triple-Negative Breast Cancer

Eunice Okome Obiang, Ph.D.
Université d'Angers in collaboration with the Institut de Cancérologie de l'Ouest, 2022

This thesis is articulated around two axes. The first one is a contribution to the study of partial Gaussian graphical models (PGGM) in high-dimensional learning. Precisely, we are interested in the multiple-output modeling, where we aim at estimating, on the one hand the matrix Δ of direct links between predictors and responses, and on the other hand the conditional precision matrix Ω_y of responses. We start with a frequentist approach by penalized maximum likelihood, where we propose a PGGM with two forms of penalization: a ℓ_1 penalty inducing sparsity on Δ and Ω_y, and a structural penalty reflecting a generalized Gaussian prior on the direct links. We show that, when properly regularized, this model comes with a theoretical guarantee taking the form of an upper bound on the estimation error. Finally, we close this first reflection with empirical studies highlighting the structuring property of this estimation procedure, and its relevance on a real dataset. We continue with the study of the Bayesian counterpart, previously unexplored in the literature. Following a spike and slab strategy, we offer several hierarchical structures imposing either a saturated, sparse, group-sparse or sparse-group-sparse configuration of the matrix Δ. We obtain a theoretical guarantee for the sparse and group-sparse configurations, and illustrate the competitive results of these models on a simulation study and a real dataset, conducted with Gibbs samplers. The second part of the thesis is entirely devoted to the selection of prognostic variables in multi-omics survival analysis. We propose a stepwise variable selection algorithm offering a consensus between different regularization methods, including those presented in the first axis. The efficiency of this approach is finally studied on a dataset relating to triple negative breast cancer, while taking care to meet the constraints identified by oncologists. All our codes are made available to the community.

Remerciements

Me voilà au terme de mes trois années de doctorat, et l'écriture des remerciements constitue la dernière épreuve à laquelle je suis confrontée. J'espère qu'en ces quelques lignes je saurais transmettre toute la gratitude que j'éprouve envers les belles personnes qui ont contribué, de près, comme de loin, au bon déroulement de cette thèse.

C'est tout naturellement que mes premiers remerciements s'adressent à Frédéric Proïa. Tu as été un mentor pour moi dès ma deuxième année de Licence, où tu m'as transmis ta passion des Statistiques. Depuis, tu m'as accompagné tout au long de mon parcours universitaire, jusqu'à ce jour. Merci de m'avoir offert l'opportunité d'effectuer une thèse, et surtout merci pour la qualité de l'encadrement que tu as fourni. En particulier, si notre collaboration s'est déroulée dans de bonnes conditions, c'est parce que tu as été aussi rigoureux et exigeant, que patient et bienveillant à mon égard. Je sors de cette expérience grandie en tant que chercheuse, c'est en grande partie grâce à ton soutien.

Je remercie bien évidemment mon directeur de thèse, Loïc Chaumont, pour la confiance qu'il m'a accordée et la bienveillance qui l'anime. Merci également à mon encadrant Pascal Jézéquel et sa collaboratrice Fadoua Ben Azzouz pour les moments d'échange autour des problématiques appliquées au cancer du sein. Plus généralement, je remercie toute l'équipe EPICURE de l'ICO, sans qui ce projet n'aurait pas vu le jour. Je vous souhaite beaucoup de réussite dans la suite de cette noble lutte contre le cancer.

Je me tourne désormais vers Julien Chiquet et Pierre Pudlo qui ont accepté de rapporter sur cette thèse. Merci énormément pour votre travail de relecture et vos rapports bienveillants sur mes travaux. Excusez-moi d'avoir ajouté une charge supplémentaire à votre rentrée académique, j'espère que cette lecture vous aura tout de même été agréable. J'exprime également toute ma gratitude à Malgorzata Bogdan, Agnès Lagnoux et Gilles Stupfler de m'avoir fait l'honneur d'être membres de mon jury en qualité d'examinatrice·eur·s. Si les circonstances ne nous permettent pas de tou·te·s nous rencontrer, j'espère toutefois que mes remerciements vous parviendront.

Je souhaite, en outre, remercier l'Université d'Angers, et plus particulièrement les membres du LAREMA avec lesquels j'ai partagé cette folle expérience qu'est le doctorat. Ayant effectué l'intégralité de mon cursus universitaire à Angers, je n'ai pas été dépaysée lors de mon arrivée au LAREMA. En revanche, j'ai apprécié découvrir d'autres facettes de mes ancien·e·s enseignant·e·s, devenu·e·s des collègues. Je repense, notamment, à la proposition de changement du nom du laboratoire en vue de faire honneur à une mathématicienne renommée. Si ce projet n'a pas abouti, il m'a fait prendre conscience que même dans cet environnement trop masculin que sont les mathématiques, la cause féministe ne laisse pas tout le monde indifférent. C'est donc chaleureusement que je souhaite remercier celleux qui ont soutenu ce noble projet, d'abord pour le projet en lui-même, mais également pour le message d'espoir qu'iels m'ont transmis. J'aimerais également remercier mes camarades de front, les doctorant·e·s du LAREMA, de celleux qui m'ont accueilli le premier jour, à celleux que je quitterai en ce dernier jour. Il est vrai que nous ne sommes pas tou·te·s très proches, mais nous sommes connecté·e·s par un horrible fléau : le syndrôme de

l'imposteur·e. En parlant de ce fléau particulièrement observé auprès des jeunes chercheu·r·se·s, j'aimerais remercier d'autres expert·e·s en la matière, les doctorant·e·s du CJB. Merci de m'avoir accueillie comme l'une des vôtres, bien que mes connaissances juridiques sont aussi maigres que celles d'un privatiste. Enfin, je remercie mes étudiant·e·s de m'avoir enseigné qu'il va de soi que rien ne va de soi.

Cela étant dit, il me tient à cœur d'offrir des remerciements appropriés aux collègues et ami·e·s qui, à point nommé, ont su trouver les mots pour me remettre d'aplomb quand tout allait de travers. D'abord j'aimerais remercier notre bonne fée Alexandra. Merci pour toutes nos belles conversations sans fin et pour ton professionnalisme dans la gestion des différentes démarches administratives qui encombrent le doctorat. Merci également à Eric, Daniel, François, Maxime, Axel, Aurore, Thomas, Thibault et Jean-Baptiste pour les pauses café souvent plus longues que prévues, et les diverses conversations qui les accompagnaient. Merci à Antoine et Sinan pour les escapades au Joker's et au Welsh. Merci à Agathe, Thomas, Alexis(ssss), Sabrina, Chloé, Maël, Florian, Pierre et Clara pour les voyages dans le temps ou ailleurs. Enfin, merci Denis pour ta joie de vivre étonnamment contagieuse.

A présent, je me tourne vers ces personnes chères à mon cœur. Ces personnes qui, sans toujours s'en rendre compte, on eu un impact considérable sur ma santé mentale, et par continuité sur le bon déroulement de mon doctorat. A commencer par mes soeurs de cœur Séta et Nora. Que dire ? Vous avez été là depuis tellement d'années. Je ne saurai vous remercier assez pour tout l'amour que vous m'avez témoigné, alors je tente le coup en vous disant simplement merci d'exister. Je remercie également Nadia et Mariam, mes jumelles préférées avec qui je me suis lancée dans cette aventure. La distance, la covid ou encore les différents CSI n'auront pas eu raison de notre trio, les camarades de galère. Merci Clément pour ton cœur tendre, que tu échoues à dissimuler malgré bien des efforts. Merci Joseph et Anaïs pour votre accueil *so British* dans le Layon. Quant à vos suggestions de séries, sachez qu'elles sont perfectibles. Merci Olivia de m'avoir appris les bruits d'animaux, à seulement 1 an tu faisais preuve d'une pédagogie sans faille. Merci camarade Ouriel d'avoir renoncé à faire de moi une adepte du bridge, au profit de conversations politiques autour d'une pinte de cidre. Merci Sofia pour ton soutien indéfectible, malgré la distance qui n'a fait que consolider notre amitié (mais ne nous éloignons pas trop non plus). Merci Théo pour les balades à moto, je tâcherai de ne pas transformer celles à venir en session de dépannage. Merci Renata et Sasha pour les meeting au 17:45, bien après 17h45, où nos différentes cultures s'harmonisaient autour de deux essentiels. Merci Elodie pour ta joie de vivre et ta positivité. Tu t'es attachée à me redonner confiance en l'humanité, j'essaierai d'être coopérative. Enfin, je remercie le milieu militant d'avoir mis sur mon chemin Kim, Audrey, Maelle, Colin et Greg. Vous m'avez offert un environnement safe, respectueux et bienveillant, et m'apprenez, chaque jour, à l'être à mon tour, merci énormément.

Je ne peux envisager cette série de remerciements sans un mot pour ma famille. Celleux dont le soutien et l'amour me semblent si acquis, que j'en oublie, trop souvent, de les remercier proprement. A commencer par mon frère Joëd, mon allié infatigable. On a beau dire que le hasard fait bien les choses, dans notre cas rien n'est moins vrai, puisqu'il a failli à faire de nous les jum·eau·elle que nous devrions être. Merci d'avoir été, tout au long de ma vie, un coach, un supporter, et un soutien (un clown ?). Je remercie également ma mère, Marie-Paulette, pour tous les sacrifices qu'elle a faits dans sa vie personnelle pour que j'en arrive là aujourd'hui. Il est plus que temps que tu penses à toi, et que tu profites dignement de ta retraite. Merci Estelle et Elliott d'avoir été pour moi, une famille, en toute simplicité. Merci aux

familles Fournier et Masson pour leur accueil chaleureux, depuis le premier jour ; vous n'êtes pas des secondes familles, vous êtes les belles. Enfin, je remercie mes frères, mes soeurs, et leurs conjoint·e·s, qui participent, malgré la distance, à mon petit bonheur quotidien.

« Et si on concluait ? » Eh bien j'aimerais conclure par quelques mots à l'égard de mon cher conjoint, Timothée Masson. Cela fait déjà plus de sept ans que nous évoluons ensemble, et si les années parlent d'elles-même, il me faut toutefois marquer le coup, en t'adressant des remerciements en bonne et due forme. Ainsi, je te remercie, Timothée, pour l'ensemble de ton œuvre. Merci d'avoir été à mes côtés pour célébrer mes victoires, et d'y être resté lorsque les jours n'avaient plus rien de victorieux. Merci d'avoir su être patient, et à l'écoute, en toutes circonstances. Merci de m'avoir appris à m'aimer, à me respecter, à me pardonner. Pour finir, merci pour ton souffle, ta joie de vivre ; ta beauté, tout simplement.

Je clos ces remerciements par un hommage à toutes les personnes qui se battent dans le monde pour l'obtention de droits fondamentaux : la Liberté et l'Egalité.

Solidarité avec les femmes, du monde entier.

Solidarité avec les trans, du monde entier.

Solidarité avec les non binaires, du monde entier.

Solidarité avec les intersexes, du monde entier.

Solidarité avec les genderfluids, du monde entier.

Solidarité avec les lesbiennes, du monde entier.

Solidarité avec les gays, du monde entier.

Solidarité avec les bis, du monde entier.

Solidarité avec les pans, du monde entier.

Solidarité avec les polys, du monde entier.

Solidarité avec les queers, du monde entier.

Solidarité avec les handi·e·s, du monde entier.

Solidarité avec les racisé·e·s, du monde entier.

Solidarité avec les discriminé·e·s, du monde entier.

Solidarité avec les ostracisé·e·s, du monde entier.

Solidarité avec les exploité·e·s, du monde entier.

"For to be free is not merely to cast off one's chains, but to live in a way that respects and enhances the freedom of others."

— Nelson Mandela

Contents

List of Figures

List of Tables

List of Algorithms

Acronyms

Acronym	Definition
CNA	Copy-Number Alteration
DAG	Directed acyclic graph
FUSCC	Fudan University Shanghai Cancer Center
GGM	Gaussian graphical model
HER2	Human Epidermal Growth Factor de type 2
NODE	National Omics Data Encyclopedia
RNAseq	RNA sequencing
PGGM	partial Gaussian graphical model
SVSSA	Stepwise Variable Selection for Survival Analysis
TNBC	Triple-Negative Breast Cancer
UG	Undirected graph

NOTATIONS

Notation	Meaning
$0_{n,m}$	null matrix of dimension $n \times m$
$\det(A)$	determinant of the square matrix A
$\lambda_i(A)$	the i-th eigenvalues of A
$\mathrm{sp}(A)$	spectrum of matrix A taken in decreasing order (from $\lambda_1(A) = \lambda_{\max}(A)$ to $\lambda_d(A) = \lambda_{\min}(A)$)
$\mathbb{S}_+^d$	cone of symmetric positive semi-definite matrices of dimension d
$\mathbb{S}_{++}^d$	cone of symmetric positive definite matrices of dimension d
$\mathrm{vec}(A)$	the vectorization of A into a column vector
$\lvert A\rvert_*$	$= \lVert\mathrm{vec}(A)\rVert_*$ the elementwise ℓ_* norm of A
$\lvert A\rvert_*^-$	$\lvert A\rvert_*$ deprived of the diagonal terms of A
$\lVert A\rVert_F$	$= \lvert A\rvert_2$ the Frobenius norm of A
$\lVert A\rVert_2$	the spectral norm of A
$\langle\!\langle A, B\rangle\!\rangle$	$= \langle \mathrm{vec}(A), \mathrm{vec}(B)\rangle = \mathrm{tr}(A^t\, B)$ the Frobenius inner product
	between any matrices A and B of same dimensions
$\langle u, v\rangle$	$= u^t\, v$ the inner product of the Euclidean real space
$\lvert u\rvert_0$	the number of non-zero values in the vector u
$A_{\setminus j}$	the square matrix A with row and column j removed
A_j^-	the column j of the square matrix A without the diagonal element A_{jj}
$[A]_C$	the matrix A whose elements outside of the set of coordinates C are set to zero
$\mathrm{sign}(x)$	the sign of the element x
$\mathrm{rank}(v)$	rank of each element of the vector v

INTRODUCTION (FRANÇAIS)

LE CANCER DU SEIN est la forme de cancer la plus répandue et la plus meurtrière chez les femmes [FCS+21]. Depuis les années 1990, de nombreux travaux de recherche ont permis l'amélioration des techniques de dépistage et le développement de thérapies plus efficaces. La mortalité du cancer du sein a ainsi été endiguée, jusqu'à diminuer progressivement, mais le taux d'incidence poursuit son augmentation globale. C'est ainsi qu'en 2018, nous comptions en France métropolitaine, quatrième pays le plus touché par cette pathologie, 58 500 nouveaux cas et 12 146 décès [dC, Int]. La lutte contre le cancer du sein demeure donc un enjeu majeur de santé publique.
Il existe plusieurs types de cancers du sein définis par des facteurs histopronostiques qui influent sur le choix du traitement et le pronostic de la maladie [dCdl]. Le cancer du sein triple négatif (TNBC) est une forme tumorale particulièrement éprouvante en oncologie. Il s'agit d'une des formes les plus agressives du cancer du sein, que l'on retrouve dans environ 15% des cas, dont majoritairement des femmes non ménopausées. Elle se caractérise par un développement et une propagation rapide de la tumeur, ainsi qu'un risque de rechute des plus élevés donnant souvent lieu à des métastases. Ses particularités font du TNBC un type de cancer difficile à pronostiquer [Soc].

Afin d'améliorer la planification de la prévention et la conception de nouveaux traitements individualisés, une meilleure compréhension des mécanismes pathologiques de l'apparition et de la progression du TNBC est nécessaire. A cet effet, de nombreuses études sont effectuées afin de déterminer les facteurs biologiques ayant une réelle influence sur le pronostic du TNBC [QXH+16, TCJL+10, DTP+07]. Les progrès scientifiques et technologiques des deux dernières décennies, en termes de collecte et de traitement de données, ont grandement impacté les méthodes de recherche des facteurs pronostiques. Les différentes unités de recherche ne se limitent plus aux données issues de leurs services respectifs, mais disposent désormais d'un large panel de données biologiques, de types, de formats et de sources divers, offrant de nouvelles perspectives de recherches cliniques. Cette dynamique a entraîné le développement d'études multi-omiques, où les praticiens exploitent plusieurs types de données moléculaires de haute dimension (telles que les données génomiques pour l'ADN, les transcriptomiques pour l'ARN, les protéomiques pour les protéines, etc.), dans le but de comprendre la biologie sous-jacente de maladies complexes comme le cancer. En revanche, si les applications sont en expansion, la recherche statistique sur l'intégration des données multi-omiques au sein d'algorithmes de *Machine Learning* peut encore gagner en maturité. Pour y voir plus clair, tournons-nous vers les défis liés à ce type de données.

D'abord rappelons que les données mutli-omiques s'inscrivent dans le cadre de la grande dimension, où le nombre d'individus est largement inférieur au nombre de variables explicatives. En particulier, les données exploitées proviennent généralement d'enquêtes de cohortes comportant quelques centaines d'individus, sur lesquels les informations collectées constituent des centaines de milliers de variables, selon les matrices omiques considérées. Ainsi, outre la nécessité de construire des modèles prédictifs offrant une

précision à la hauteur des enjeux sanitaires, il est essentiel de considérer également le caractère pratique de ces modèles. Un modèle construit sur la base de milliers de variables, mobilisant de nombreuses procédures de collecte, aura une utilité pratique limitée face à un modèle de plus petite taille. La recherche de la parcimonie, et donc la sélection de variable constitue le premier enjeu de ces études cliniques. La littérature statistique renferme un large éventail de méthodes de sélection de variables pour l'analyse en grande dimension. La plus connue, et très saluée par la littérature est la méthode *Lasso* de Tibshirani [Tib96]. Considérons un problème de régression linéaire, où nous souhaitons exprimer une variable $Y \in \mathbb{R}$ à partir d'un vecteur de p variables explicatives $X \in \mathbb{R}^p$, par la forme $Y = \beta^t X + \epsilon$ où β est le vecteur des coefficients de régression et ϵ est un bruit. En s'appuyant sur un échantillon de n observation indépendantes $(Y_i, X_i)_{i=1,..,n}$, la méthode *Lasso* résout le problème de régression pénalisé par la norme ℓ_1 suivant

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big(\sum_{i=1}^{n} (Y_i - \beta^t X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big), \tag{1}$$

où λ est un paramètre de pénalisation. Ce problème est équivalent à la minimisation de l'erreur quadratique sous une contrainte de la forme $|\beta|_1 \leqslant s$, où s est un paramètre de régularisation. Ainsi défini, la méthode *Lasso* a des vertus à la fois de *shrinkage* et de parcimonie, et plus précisément l'estimateur obtenu a un support $S = \{j; \hat{\beta}_j \neq 0\}$ vérifiant $\text{card}(S) \leqslant n$. Toutefois, si le *Lasso* offre de bonnes performances lorsque la taille du support du modèle oracle est effectivement inférieur à n, cette méthode ne répond pas intégralement aux contraintes liées aux données mutli-omiques. D'abord, la procédure de sélection de variables s'avère trop restrictive en raison de la taille relativement faible des échantillons issus des enquêtes de cohortes. De plus, lorsqu'il est appliqué à des groupes de variables fortement corrélées, le lasso aura tendance à sélectionner arbitrairement une seule variable du groupe, ignorant les actions conjointes des prédicteurs que l'on peut par exemple retrouver au sein des réseaux de régulation de gènes, ou encore au sein d'associations de traits phénotypiques présents dans les mesures omiques [FGFBGM22, Wai09, ZY06]. Cette limite nous mène au deuxième enjeu des études cliniques, la prise en compte de la structure de corrélation des prédicteurs dans les procédures de sélection de variables.

Cette problématique a été traitée par de nombreux auteurs, et sous différents formats des plus remarquables. Considérons à présent que les prédicteurs sont ordonnés en m groupes, tel que $X = (\underline{X}_1, ..., \underline{X}_m) \in \mathbb{R}^p$. Yuan et Lin ont étudié la méthode *Group Lasso* effectuant la sélection de variables à l'échelle des groupes, en remplaçant la pénalisation ℓ_1 du *Lasso* par une pénalisation ℓ_1-ℓ_2 sur le vecteur de coefficients associé à chaque groupe [YL06]. Cette approche fut initiée dans le cadre de la régression linéaire par la minimisation de l'erreur quadratique, et a été par la suite étendue aux fonctions de pertes générales dans l'article de Kim *et al.* [KKK06]. Précisons toutefois que l'approche de sélection par groupes a été initialement pensée pour intégrer des variables catégorielles à la régularisation ℓ_1. Ainsi, si ces méthodes sont efficaces pour le traitement de petits groupes de variables, elles offriraient un résultat peu sparse sur un groupement omique. Le recours à celles-ci doit justifier une segmentation préalable des variables au sein des groupes omiques, pouvant être réalisée selon une connaissance a priori des interactions entre les prédicteurs, ou encore par le biais d'un clustering [MSH07]. Afin de pallier l'excès de générosité du *Group Lasso*, Simon *et al.* [SFHT13] ont proposé la méthode *Sparse-Group Lasso* qui effectue la sélection de variables sur deux niveaux ; d'abord à l'échelle des groupes par la pénalisation ℓ_1-ℓ_2, puis à l'échelle des

variables au sein des groupes non nuls par la pénalisation ℓ_1. La littérature sur ce sujet est très fournie. Citons simplement les travaux de Li *et al.* [LNZ15], la régression linéaire avec spike and slab de Xu et Gosh [XG15] ainsi que la généralisation vectorielle de Liquet *et al.* [LMPS17].

La modélisation à sorties multiples constitue précisément le troisième enjeu que nous souhaitons soulever. Supposons maintenant que nous ayons affaire à une régression linéaire multivariée de la forme

$$Y = B^t X + E, \tag{2}$$

où $Y \in \mathbb{R}^q$ est le vecteur des réponses, $X \in \mathbb{R}^q$ le vecteur des variables explicatives (éventuellement structurées en groupes), $B \in \mathbb{R}^{p\times q}$ la matrice des coefficients de régression et $E \in \mathbb{R}^p$ un terme de bruit gaussien multivarié. Le modèle graphique gaussien partiel (PGGM), développé par Sohn et Kim [SK12] ou Yuan et Zhang [YZ14], apparaît comme un outil puissant pour mettre en évidence les relations entre prédicteurs et réponses par le biais de corrélations partielles (appelées désormais 'liens directs', par opposition aux 'liens indirects' résultant de corrélations). En effet, supposons que le couple $(Y, X) \in \mathbb{R}^{q+p}$ admet également une distribution gaussienne multivariée de moyenne nulle, de covariance Σ et de matrice de précision $\Omega = \Sigma^{-1}$. La décomposition en blocs donnée par

$$\Omega = \begin{pmatrix} \Omega_y & \Delta \\ \Delta^t & \Omega_x \end{pmatrix} \tag{3}$$

avec $\Omega_y \in \mathbb{S}^q_{++}$, $\Delta \in \mathbb{R}^{q\times p}$ et $\Omega_x \in \mathbb{S}^p_{++}$ conduit à $Y_k \,|\, X_k \sim \mathcal{N}_q(-\Omega_y^{-1}\Delta\, X_k, \Omega_y^{-1})$. Cette remarque est cruciale car on peut voir que la régression à sorties multiples $Y_k = B^t X_k + E_k$ avec un bruit gaussien $E_k \sim \mathcal{N}_q(0, R)$ peut être reparamétrée avec

$$B = -\Delta^t \Omega_y^{-1} \quad \text{et} \quad R = \Omega_y^{-1}. \tag{4}$$

Ainsi, tandis que Δ ne contient que les liens directs entre les prédicteurs et les réponses, l'estimation de B est impactée par la structure de corrélation des sorties. Afin d'assurer une sélection de variables pertinentes, Yuan et Zhang [YZ14] proposent d'approcher B à travers une estimation distincte de Δ et Ω_y^{-1} par maximum de vraisemblance pénalisée induisant de la sparsité au sein des matrices Δ et Ω_y^{-1}. Cette approche fut ensuite reprise par Chiquet *et al.* [CMHR17]. Les auteurs proposent une variante dans laquelle aucune hypothèse de sparsité n'est portée sur Ω_y, et centrent leur schéma de régularisation sur les liens directs Δ. En particulier, celui-ci prend la forme de deux pénalisations, une pénalisation ℓ_1 d'une part, et une pénalisation structurante d'autre part induite par une connaissance a priori de la structure de corrélation des prédicteurs.

Avec ce panorama en tête, cette thèse s'est articulée autour de deux axes. Le premier axe, théorique, apporte des contributions à la théorie des modèles graphiques gaussiens partiels. Notre première étude reprend les travaux de Yuan et Zhang [YZ14] et ceux de Chiquet *et al.* [CMHR17]. En particulier nous proposons une méthode d'estimation et de sélection de variables, où nous appliquons une pénalisation de type ℓ_1 sur Δ et Ω_y afin d'introduire de la sparsité, et une pénalisation structurante sur Δ reflétant un a priori gaussien généralisé sur les liens directs. Nous apportons une garantie théorique à notre méthode,

et la mettons en application sur des données synthétiques et un ensemble de données réelles sont menées. Inspirés par les travaux de Xu et Gosh [XG15], et Liquet *et al.* [LMPS17] traitant l'estimation bayésienne de B, notre deuxième étude explore diverses approches bayésiennes au sein des modèles graphiques gaussiens partiels. En suivant la stratégie spike and slab de ces auteurs, nous proposons différents modèles hiérarchiques permettant de traiter des régressions linéaires à sorties uni ou multidimensionnelles, en petite ou grande dimension, selon une configuration saturée, sparse, group sparse ou encore sparse-group-sparse de la matrice Δ. Nous fournissons également une garantie théorique sur nos estimations sparse et group-sparse, et montrons l'efficacité de nos modèles d'abord à travers des différents scénarios de données simulées et puis par l'étude d'un jeu de données réel. Les résultats obtenus sont très compétitifs, notamment en termes de récupération de support.

Le deuxième axe de la thèse a un objectif plus appliqué à l'analyse de survie, mais s'intègre également dans la problématique de sélection de variables. Nous nous situons dans le cadre d'une analyse multi-omique où nous souhaitons exprimer la survie de patientes à partir d'un échantillon d'observations contenant de la censure. Les méthodes de sélection de variables en analyse de survie ont été majoritairement été étudiée dans le cadre des modèles de Cox, traitant différents enjeux liés à l'analyse multi-omique. Nous pensons notamment à la méthode priority-Lasso de Klau *et al.* [KJH+18] qui tient compte d'une connaissance a priori de l'usabilité des groupes de variables en pratique, ou encore l'IPF-Lasso de Boulesteix *et al.* [BDBJF17] qui attribue différents facteurs de pénalisation aux groupes afin de contrebalancer le déséquilibre induit par les différences de tailles des groupes. Néanmoins, si le modèle de Cox est de loin le plus répandu dans ces études [Cox75], les modèles de Cure mériteraient une attention particulière, puisqu'ils sont spécifiquement conçus pour traiter les problématique cliniques où l'on peut vraissemblablement considérer qu'une partie de la population censurée est tout simplement immunisée et n'observera pas l'évènement [Boa49]. Cependant, la littérature compte peu d'études de sélection de variable reprenant cette approche.
Nous proposons un algorithme de sélection de variables descendante nommée Stepwise Variable Selection for Survival Analysis (SVSSA) dans ce manuscrit. L'idée est d'assurer la fiabilité de la sélection d'une part en la segmentant en quatre étapes et d'autre part en effectuant un consensus entre des méthodes de régularisation bien établies dans la littérature. En particulier, nous supposons qu'une variable définie comme pertinente par différents modèles de sélections, reposant sur des critères et des approches variés, est vraisemblablement significative. Nous testons cette procédure sur un jeu de données portant sur le cancer du sein triple négatif, en comparant les performances des modèles établis dans la littérature à celles des modèles de Cox et de Cure construits à partir des variables sélectionnées par SVSSA). Les résultats sont très compétitifs, mais la méthode gagnerait en optimisation du temps de calcul.

Organisation du manuscrit

Ce manuscrit s'articule principalement autour de deux axes. Le premier axe, théorique, est composé des trois premiers chapitres dédiés aux modèles graphiques gaussiens. Le deuxième axe, appliqué, se compose quant à lui des deux derniers chapitres, et traite de la sélection de variables en analyse de survie en grande dimension. Plus précisément :

(1) Le chapitre 1 offre une introduction succincte à la théorie des graphes, et en particulier aux modèles

graphiques gaussiens. Nous commençons par définir les concepts élémentaires de la théorie des graphes, et mettons en lumière la relation entre la structure d'indépendances conditionnelles des variables et la matrice de précision du modèle graphique. Nous poursuivons par une présentation des modèles graphiques gaussiens et leurs avantages dans la récupération des liens directs entre prédicteurs et les réponses par rapport aux modèles de régression linéaire standards. Ensuite, nous présentons le modèle graphique gaussien partiel pénalisé introduit par Sohn et Kim [SK12] et Yuan et Zhang [YZ14], ainsi que ses avantages en termes de précision et de temps de calcul. Enfin, nous discutons de la littérature bayésienne des modèles graphiques gaussiens en grande dimension.

(2) Le chapitre 2 présente les résultats obtenus dans le premier article de cette thèse, ayant fait l'objet d'une publication dans le journal *ESAIM: Probability and Statistics* [OOJP21]. Il est consacré à l'estimation d'un modèle graphique gaussien partiel avec une pénalisation structurelle sur les liens directs. Après une présentation de la littérature qui aborde cette problématique, nous formalisons l'écriture de notre pénalisation gaussienne généralisée et exposons la nouvelle fonction objectif de notre PGGM. Nous poursuivons sur les garanties théoriques de notre estimateur, et détaillons la preuve du théorème principal. Enfin, nous testons les performances de notre méthode d'abord par des études empiriques sur des données synthétiques, puis en traitant un jeu de données réel, avant de conclure sur les perspectives d'évolution.

(3) Le chapitre 3 présente les résultats obtenus dans le deuxième article de cette thèse, ayant fait l'objet d'une publication dans le journal *Bayesian Analysis* [OOJP22]. Après une présentation des travaux qui nous ont précédés dans la littérature, nous formalisons l'écriture de nous modèles hiérarchiques. Nous poursuivons sur les garanties théoriques associées aux configuration sparse et group-sparse. Puis nous testons nos méthodes sur des échantillonneurs de Gibbs et un jeu de données réel, avant de conclure.

(4) Le chapitre 4 fait office d'introduction à l'analyse de survie. Nous commençons par définir le cadre et les problématiques spécifiques aux modèles de survie, et poursuivons sur les principales méthodes d'estimation de tels modèles dans le cadre non paramétrique, semi-paramétrique et paramétrique. Nous abordons ensuite la question de la sélection de variables en survie, au travers des méthodes d'estimation pénalisées, et concluons sur les modèles de Cure et leurs spécificités pour la modélisation d'évènements cliniques.

(5) Le chapitre 5 présente les derniers résultats de cette thèse, à savoir une méthode de sélection de variables descendante adaptée aux études multi-omiques. La première partie du chapitre s'attelle à détailler le fonctionnement de notre algorithme, en apportant notamment une description succincte des différentes méthodes qui le composent. La deuxième partie est une mise en pratique sur des données réelles portant sur le cancer du sein triple négatif, dans laquelle nous comparons les résultats de notre approche avec des méthodes de régularisation bien établies, selon la pertinence de la sélection, les performances prédictives, et le temps de calcul.

Introduction

Breast cancer is the most common and deadliest form of cancer in women [FCS+21]. Since the 1990s, numerous research studies have led to improved screening techniques and the development of more effective therapies. Mortality from breast cancer has been contained and is gradually decreasing, but the overall incidence rate continues to rise. France ranks fourth worldwide by incidence of breast cancer. In 2018, 58 500 new cases and 12 146 deaths were recorded in Metropolitan France alone [dC, Int]. The fight against breast cancer therefore remains a major public health issue.
There are several types of breast cancer defined by histopronostic factors that influence the choice of treatment and the prognosis of the disease [dCdl]. Triple-negative breast cancer (TNBC) is a particularly challenging tumor form in oncology. It is one of the most aggressive forms of breast cancer, occurring in approximately 15% of cases, mostly in premenopausal women. It is characterized by a rapid development and spread of the tumor, as well as a high risk of relapse, often resulting in metastasis. Its particularities make TNBC a type of cancer difficult to prognose [Soc].

In order to improve prevention planning and the design of new individualized treatments, a better understanding of pathological mechanisms of TNBC onset and progression is needed. To this end, numerous studies are being performed to determine the biological factors that actually influence the TNBC prognosis [QXH+16, TCJL+10, DTP+07]. Scientific and technological advances over the last two decades, in terms of data collection and processing, have greatly impacted the methods of research of prognostic factors. The various research units are no longer limited to data from their respective departments, they now have a large panel of biological data, of various types, formats and sources, offering new perspectives for clinical research. This dynamic has led to the development of multi-omics studies, where practitioners exploit several types of high-dimensional molecular data (such as genomic data for DNA, transcriptomic data for RNA, proteomic data for proteins, etc.), in order to understand the underlying biology of complex diseases such as cancer. On the other hand, even though its applications are expanding, statistical research on the integration of multi-omics data within machine learning algorithms is still far from being a mature field. To get a clearer picture, let us turn to the challenges related to this type of data.

First of all, let us recall that mutli-omics data are part of the high-dimensional framework, where the number of individuals is much lower than the number of explanatory variables. In particular, the data used generally comes from cohort surveys comprising a few hundred individuals, on which the information collected constitutes hundreds of thousands of variables, depending on the omic matrices considered. Thus, in addition to the need to build predictive models offering an accuracy commensurate with the health issues at stake, it is essential to also consider the practicality of these models. A model built with thousands of variables, mobilizing many collection procedures, will have limited practical utility compared to a smaller model. The search for sparsity, and therefore the selection of variables, constitutes the first challenge of these clinical studies. The statistical literature contains a wide range of variable

selection methods for high-dimensional analysis. The best known, and highly acclaimed in the literature, is the Lasso method of Tibshirani [Tib96]. Consider a linear regression problem, where we wish to express a variable $Y \in \mathbb{R}$ from a vector of p explanatory variables $X \in \mathbb{R}^p$, by the form $Y = \beta^t X + \epsilon$ where β is the vector of regression coefficients and ϵ is a random noise. Based on a sample of n independent observations $(Y_i, X_i)_{i=1,..,n}$, the Lasso method solves the following ℓ_1 penalized regression problem

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \Big(\sum_{i=1}^{n} (Y_i - \beta^t X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big), \tag{1}$$

where λ is a penalty parameter. This problem is equivalent to minimizing the sum of squared errors under a constraint of the form $|\beta|_1 \leqslant s$, where s is a regularization parameter. Thus defined, the Lasso method has both shrinkage and sparsity virtues, and more precisely the estimator obtained has a support $S = \{j; \hat{\beta}_j \neq 0\}$ verifying $\text{card}(S) \leqslant n$. However, while Lasso performs well when the support size of the oracle model is indeed less than n, this method does not fully meet the constraints related to mutli-omic data. First, the variable selection procedure turns out to be too restrictive due to the relatively small size of the samples from cohort surveys. Moreover, when applied to groups of highly correlated variables, the Lasso will tend to arbitrarily select a single variable from the group, ignoring the joint actions of predictors that can be found, for example, in gene regulation networks, or in associations of phenotypic traits present in omic measures [FGFBGM22, Wai09, ZY06]. This limitation leads us to the second challenge of clinical studies, taking into account the correlation structure of the predictors in variable selection procedures.

This problem has been treated by many authors, and in various remarkable formats. Suppose we have omic predictors which are arranged into m groups, such that $X = (\underset{\sim}{X}_1, ..., \underset{\sim}{X}_m) \in \mathbb{R}^p$. Yuan and Lin studied the Group Lasso method, which performs group-wide variable selection, replacing the Lasso ℓ_1 penalty by a ℓ_1-ℓ_2 penalty on the vector of coefficients associated with each group [YL06]. This approach was initiated in the context of linear regression by minimizing the sum of squared errors, and was later extended to general loss functions in the paper by Kim *et al.* [KKK06]. Note however that the group selection approach was initially designed to integrate categorical variables into the ℓ_1-regularization. Thus, although these methods are efficient at dealing with small groups of variables, they would not offer a result sparse enough on an omic grouping. The use of these methods must justify a preliminary segmentation of the variables within the omic groups, which can be carried out according to prior knowledge of the interactions between predictors, or even by means of a clustering [MSH07]. In order to compensate the excessive generosity of the Group Lasso, Simon *et al.* [SFHT13] proposed the method Sparse-Group Lasso which performs variable selection on two levels ; first at the group level with the ℓ_1-ℓ_2 penalty, then at the variable level within the non-zero groups with the ℓ_1 penalty. The literature on this subject is extensive. In particular we note the works of Li *et al.* [LNZ15], the linear regression with spike and slab of Xu and Gosh [XG15] as well as the vector generalization of Liquet *et al.* [LMPS17].

The third challenge we want to raise is the need for multi-output modeling. Suppose now that we deal with a multivariate linear regression of the form

$$Y = B^t X + E, \tag{2}$$

where $Y \in \mathbb{R}^q$ is the vector of responses, $X \in \mathbb{R}^q$ the vector of predictors (possibly structured in groups), $B \in \mathbb{R}^{p\times q}$ the matrix of regression coefficients and $E \in \mathbb{R}^p$ a multivariate Gaussian noise term. The partial Gaussian graphical model (PGGM), developped *e.g.* by Sohn and Kim [SK12] or Yuan and Zhang [YZ14], appears as a powerful tool to exhibit relationships between predictors and responses that exist through partial correlations (called from now on 'direct links', as opposed to 'indirect links' resulting from correlations). Indeed, assume that the couple $(Y, X) \in \mathbb{R}^{q+p}$ also admits a multivariate gaussian distribution with zero mean, covariance Σ and precision matrix $\Omega = \ Sigma^{-1}$. Then, the block decomposition given by

$$\Omega = \begin{pmatrix} \Omega_y & \Delta \\ \Delta^t & \Omega_x \end{pmatrix} \tag{3}$$

with $\Omega_y \in \mathbb{S}^q_{++}$, $\Delta \in \mathbb{R}^{q\times p}$ and $\Omega_x \in \mathbb{S}^p_{++}$ leads to $Y_i \,|\, X_i \sim \mathcal{N}_q(-\Omega_y^{-1}\Delta\, X_i, \Omega_y^{-1})$. This is a crucial remark because one can see that the multiple-output regression $Y_i = B^t\, X_i + E_{\cdot i}$ with Gaussian noise $E_k \sim \mathcal{N}_q(0, R)$ may be reparametrized with

$$B = -\Delta^t\, \Omega_y^{-1} \quad \text{and} \quad R = \Omega_y^{-1}. \tag{4}$$

Thus, while Δ only contains the direct links between the predictors and the responses, the estimation of B is impacted by the correlation structure of the outputs. In order to ensure a selection of relevant variables, Yuan and Zhang [YZ14] propose to approximate B through distinct estimations of Δ and Ω_y^{-1} by penalized maximum likelihood inducing sparsity within the matrices Δ and Ω_y^{-1}. This approach was then taken up by Chiquet *et al.* [CMHR17]. The authors propose a variant in which no sparsity assumption is made on Ω_y, and focus their regularization scheme on the direct links Δ. In particular, the latter takes the form of two penalties, a ℓ_1 penalty on the one hand, and a structured penalty on the other hand induced by a prior knowledge of the correlation structure of the predictors.

With this panorama in mind, this thesis is articulated around two axes. The first axis, theoretical, brings contributions to the theory of partial Gaussian graphical models. Our first study takes up the work of Yuan and Zhang [YZ14] and that of Chiquet *et al.* [CMHR17]. In particular, we propose a method of estimation and variable selection, where we apply ℓ_1-type penalty on Δ and Ω_y to introduce sparsity, as well as structural penalty on Δ reflecting a generalized Gaussian prior on the direct links. We bring theoretical guarantees to our method, and test it on synthetic and real datasets.
Inspired by the work of Xu and Gosh [XG15], and Liquet *et al.* [LMPS17] dealing with Bayesian estimation of B, our second study explores various Bayesian approaches within partial Gaussian graphical models. Following the spike and slab strategy of these authors, we propose different hierarchical models allowing to process linear regressions with uni or multidimensional outputs, in small or high dimension, according to a saturated, sparse, group-sparse or even sparse-group-sparse configuration of the matrix Δ. We also provide a theoretical guarantee on our sparse and group-sparse settings, and show the efficiency of our models first through different simulated data scenarios and then by studying a real dataset. The results obtained are very competitive, especially in terms of support recovery.

The second axis of the thesis is related to survival analysis, but is also linked to the problem of variable selection. We are in the context of a multi-omics analysis where we want to express patient

survival times from a sample containing censored observations. Variable selection methods in survival analysis have been mainly studied in the context of Cox models, dealing with different issues related to multi-omics analysis. We think in particular of the priority-Lasso method of Klau *et al.* [KJH+18] which takes into account a usability criterion. In addition to seeking sparsity, it also prioritize selecting variables that are more routinely or more easily assessed. We should also mention the IPF-Lasso de Boulesteix *et al.* [BDBJF17] which assigns different penalization factors to groups in order to handle the imbalance induced by the differences in group sizes. Even though the Cox model is by far the most widespread in these studies [Cox75], the Cure models deserve special attention, since they are specifically designed to deal with clinical studies where one can assume that part of the censored population is simply immune and will not observe the event [Boa49]. However, there are few variable selection studies in the literature that take this approach.

We propose a stepwise variable selection algorithm named Stepwise Variable Selection for Survival Analysis (SVSSA) in this manuscript. The idea is to ensure the reliability of the selection firstly by segmenting it into four stages et also by making a consensus between regularization methods well established in the literature. In particular, we assume that a variable defined as relevant by different selection models, based on various criteria and approaches, is likely to be significant. We test this procedure on a Triple Negative Breast Cancer dataset, comparing the performances of the models established in the literature with those of the Cox and Cure models built from the variables selected by SVSSA. The results are very competitive, but the computation time must be improved.

Outline

This manuscript is mainly articulated around two axes. The first axis, theoretical, is composed of the first three chapters dedicated to Gaussian graphical models. The second axis, applied, is composed of the last two chapters, and deals with variable selection in high-dimensional survival analysis. More precisely:

(1) Chapter 1 provides a brief introduction to graph theory, and in particular to Gaussian graphical models. We start by defining the basic concepts of graph theory, and highlight the relationship between the structure of conditional independences of the variables and the precision matrix of the graphical model. We continue with a presentation of Gaussian graphical models and their advantages in recovering direct links between predictors and responses compared to standard linear regression models. Next, we present the penalized partial Gaussian graphical model introduced by Sohn and Kim [SK12] and Yuan and Zhang [YZ14], and its advantages in terms of accuracy and computational time. Finally, we discuss the Bayesian literature on high-dimensional Gaussian graphical models.

(2) Chapter 2 presents the results obtained in the first article of this thesis, published in the journal *ESAIM: Probability and Statistics* [OOJP21]. It is devoted to the estimation of a PGGM with a structural penalty on the direct links. After a presentation of the literature that addresses this problem, we formalize the form of our generalized Gaussian penalization and expose the new objective function of our PGGM. We continue on the theoretical guarantees of our estimator, and detail the proof of the main theorem. Finally, we test the performance of our method first by empirical studies on synthetic data, then by processing a real dataset, before concluding on possible improvements.

(3) Chapter 3 presents the results obtained in the second article of this thesis, published in the journal *Bayesian Analysis* [OOJP22]. After a presentation of the works that preceded us in the literature, we formalize our hierarchical models. We continue on the theoretical guarantees associated with sparse and group-sparse configurations. Then test our methods on Gibbs samplers and a real dataset, before concluding.

(4) Chapter 4 serves as an introduction to survival analysis. We begin by defining the framework and the specific problems of survival models, and continue on the main methods for estimating such models in the non-parametric, semi-parametric and parametric frameworks. We then address the issue of variable selection in survival analysis, through penalized estimation methods, and conclude with a discussion of Cure models and their specificities for modeling clinical events.

(5) Chapter 5 presents the last results of this thesis, namely a stepwise variable selection method suitable for multi-omics studies. The first part of the chapter focuses on detailing the architecture of our algorithm, in particular by providing a brief description of the different methods that compose it. The second part is an application on real data on triple negative breast cancer, in which we compare the results of our approach with well-established regularization methods, according to the relevance of the selection, the predictive performances, and the computation time.

Chapter 1

Introduction to Gaussian graphical models

This introductory chapter deals with Gaussian graphical models (GGM). It describes the general framework in which the methods presented in Chapters 2 and 3 are embedded. Throughout these introductory developments, we will rely mainly on the books by Lauritzen [Lau96], Whittaker [Whi09] and Maathuis *et al.* [MDLW18], to which we refer the reader for a further introduction to graph theory. We will however borrow some additional references, these will be presented in due time.

1.1 Elementary concepts

In this section, we briefly recall the basic notions needed to understand graph theory, and in particular Gaussian graphical models.

1.1.1 Conditional independence

Consider the random vector $X = (X_1, ..., X_p)$ with probability density $f_X > 0$, and set $\Gamma = \{1, ..., p\}$. For any set $A \subset \Gamma$, we denote X_A the random vector defined by $(X_i)_{i \in A}$ and f_{X_A} its probability density.

Definition 1.1.1 (Independence)**.** Let A, B be two subsets of Γ, and X_A, X_B the associated random vectors. X_A and X_B are independent if and only if

$$f_{(X_A,X_B)}(x_A, x_B) = f_{X_A}(x_A) f_{X_B}(x_B), \qquad \forall\, x_A, x_B. \tag{1}$$

This is denoted by $X_A \perp\!\!\!\perp X_B$.

Definition 1.1.2 (Conditional independence)**.** Let A, B, C be three subsets of Γ, and X_A, X_B, X_C the associated Gaussian vectors. X_A and X_B are conditionally independent on X_C if and only if

$$\begin{aligned} & f_{(X_A,X_B)|X_C}(x_A, x_B; x_C) = f_{X_A|X_C}(x_A; x_C) f_{X_B|X_C}(x_B; x_C) \\ \iff \quad & f_{X_A|X_B,X_C}(x_A; x_B, x_C) = f_{X_A|X_C}(x_A; x_C), \qquad \forall\, x_A, x_B, x_C. \end{aligned} \tag{2}$$

This is denoted by $X_A \perp\!\!\!\perp X_B | X_C$.

The conditional independence describes the fact that as soon as we know X_C, the knowledge of X_B will not bring any additional information about X_A. We will see later that this notion is fundamental in graph

theory.

1.1.2 Partial correlation

Definition 1.1.3 (Correlation)**.** Let X_a and X_b be two random variables with strictly positive variances. The linear correlation coefficient between X_a and X_b is given by

$$\rho(X_a, X_b) = \frac{\mathrm{Cov}(X_a, X_b)}{(\mathrm{Var}(X_a)\mathrm{Var}(X_b))^{\frac{1}{2}}}. \tag{3}$$

Definition 1.1.4 (Partial correlation)**.** Let X_a, X_b and X_c be three random variables with strictly positive variances. The (linear) correlation coefficient between X_a and X_b conditional on X_c is given by

$$\rho(X_a, X_b|X_c) = \frac{\mathrm{Cov}(X_a, X_b|X_c)}{(\mathrm{Var}(X_a|X_c)\mathrm{Var}(X_b|X_c))^{\frac{1}{2}}}. \tag{4}$$

The correlation coefficient expresses the degree of linear relationship between two variables, while the partial correlation coefficient measures the relationship between the two variables conditionally on others. Correlation and independence are closely related. Independence between two variables implies a zero correlation between them, but the converse is only true in certain special cases, in particular the Gaussian case.

Consider now the Gaussian random vector $X = (X_1, ..., X_p)$ with zero mean and covariance $\Sigma \in \mathbb{S}^p_{++}$, and set $\Sigma^{-1} = \Omega$ and $\Gamma = \{1, ..., p\}$. The density of X is given by

$$f_X(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{\langle x, \Omega x\rangle}{2}\right), \qquad \forall\, x \in \mathbb{R}^p. \tag{5}$$

For any set $A \subset \Gamma$, we denote X_A the Gaussian vector defined by $(X_i)_{i\in A}$, and for more clarity we will denote B the complementary subset of A in Γ, X_B being the associated Gaussian vector (*i.e.* $B = \Gamma\backslash A$ and $X_B = (X_i)_{i\in\Gamma\backslash A}$). X, Σ and Ω thus satisfy the following decompositions into blocks

$$X = \begin{pmatrix} X_A \\ X_B \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_A & \Sigma_{AB} \\ \Sigma_{AB}^t & \Sigma_B \end{pmatrix} \qquad \text{et} \qquad \Omega = \begin{pmatrix} \Omega_A & \Omega_{AB} \\ \Omega_{AB}^t & \Omega_B \end{pmatrix} \tag{6}$$

Proposition 1.1.5 (Marginal and conditional density)**.** *Let X, X_A, X_B, Σ and Ω be defined and decomposable as above,*

(a) the marginal distribution of X_A is $\mathcal{N}(0, \Sigma_A)$;

(b) the conditional probability distribution of X_A given $X_B = x_B$ is $\mathcal{N}(-\Omega_A^{-1}\Omega_{AB}x_B, \Omega_A^{-1})$.

Proof. We only focus here on the proof of *(b)*, that of *(a)* being well known. Using the Schur complement

of the decomposition (6), we can rewrite the matrix Ω as

$$\begin{aligned}\Omega &= \begin{pmatrix} (\Sigma_A - \Sigma_{AB}\Sigma_B^{-1}\Sigma_{AB}^t)^{-1} & -(\Sigma_A - \Sigma_{AB}\Sigma_B^{-1}\Sigma_{AB}^t)^{-1}\Sigma_{AB}\Sigma_B^{-1} \\ -\Sigma_B^{-1}\Sigma_{AB}^t(\Sigma_A - \Sigma_{AB}\Sigma_B^{-1}\Sigma_{AB}^t)^{-1} & \Sigma_B^{-1} + \Sigma_B^{-1}\Sigma_{AB}^t(\Sigma_A - \Sigma_{AB}\Sigma_B^{-1}\Sigma_{AB}^t)^{-1}\Sigma_{AB}\Sigma_B^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \Omega_A & -\Omega_A\, M \\ -M^t\,\Omega_A & \Sigma_B^{-1} \,+\, M^t\,\Omega_A\, M \end{pmatrix},\end{aligned}$$

with $M = -\,\Omega_A^{-1}\Omega_{AB} = \Sigma_{AB}\Sigma_B^{-1}$. For x_B fixed, the conditional density of X_A giving $X_B = x_B$ thus satisfies

$$\begin{aligned} f_{X_A|X_B}(x_A; x_B) \;&=\; \frac{f_X(x_A, x_B)}{f_{X_B}(x_B)} \\ &\propto \exp\left(-\frac{1}{2}\big((x_A, x_B)^t\,\Omega\,(x_A, x_B) - x_B^t\,\Sigma_B^{-1}\,x_B\big)\right) \\ &\propto \exp\left(-\frac{1}{2}\big(x_A^t\,\Omega_A\,x_A + x_B^t\,\Omega_{AB}^t\,x_A + x_A^t\,\Omega_{AB}\,x_B + x_B^t\,\Omega_B\,x_B - x_B^t\,\Sigma_B^{-1}\,x_B\big)\right) \\ &\propto \exp\left(-\frac{1}{2}\big(x_A^t\,\Omega_A\,x_A - 2\,x_A^t\,\Omega_A M\,x_B + x_B^t\,(\Sigma_B^{-1} \,+\, M^t\,\Omega_A\,M)\,x_B - x_B^t\,\Sigma_B^{-1}\,x_B\big)\right) \\ &\propto \exp\left(-\frac{1}{2}\big((x_A \,-\, M\,x_B)^t\,\Omega_A\,(x_A \,-\, M\,x_B) + x_B^t\,\Sigma_B^{-1}\,x_B - x_B^t\,\Sigma_B^{-1}\,x_B\big)\right) \\ &\propto \exp\left(-\frac{1}{2}\langle\,(x_A \,-\, M\,x_B),\,\Omega_A\,(x_A \,-\, M\,x_B)\,\rangle\right). \end{aligned}$$

We obtain the form of the desired density. □

Since the independence between Gaussian variables can be shown by the nullity of the correlation coefficients, and therefore of the covariances, we can deduce from proposition 1.1.5 the following corollary.

Corollary 1.1.6. *Let X, Γ, Σ and Ω defined as above, $\forall\, i, j \in \Gamma$,*

(a) $X_i \perp\!\!\!\perp X_j \iff \Sigma_{ij} = 0$ *;*

(b) $X_i \perp\!\!\!\perp X_j | X_{\Gamma\setminus\{i,j\}} \iff \Omega_{ij} = 0.$

Proof. (b) Let pose $A = \{i, j\}$. The conditional covariances can be read on Ω_A^{-1}, or Let $A = \{i, j\}$. Conditional covariances read on Ω_A^{-1}, where

$$\Omega_A^{-1} = \frac{1}{\det(\Omega_A)}\begin{pmatrix} \Omega_{jj} & -\Omega_{ij} \\ -\Omega_{ji} & \Omega_{ii} \end{pmatrix}.$$

Thus, $\mathrm{Cov}(X_i, X_j | X_{\Gamma\setminus\{i,j\}}) = 0 \iff (\Omega_A^{-1})_{12} = 0 \iff \Omega_{ij} = 0.$ □

1.1.3 Graph - notations and terminology

Definition 1.1.7 (Graph)**.** A graph $\mathcal{G}$ is a pair $\mathcal{G} = (\Gamma, E)$, where Γ is a finite set of elements called vertices or nodes, and E is a set of edges composed of pairs of elements taken from Γ.

A graph $\mathcal{G}$ is therefore a set of nodes connected by edges. However, there are several types of graphs defined by the characteristics of their vertices and/or their edges. In this chapter, we will limit ourselves to so-called simple graphs, which are characterized by the absence of loops and multiple edges. In other words, these are graphs that have no more than one edge between a pair of vertices and in which the edges start and end on different vertices. To go further, we refer the interested reader to the beautiful introductions to graph theory offered by the books of [W+01] and [GYA18].

Definitions 1.1.8 (Undirected edge and graph)**.** Let $\mathcal{G} = (\Gamma, E)$ be a graph,

(i) An edge (i, j) of $\mathcal{G}$ is said to be undirected if $(i, j) \in E$ and $(j, i) \in E$. It is then denoted by $i, \leftrightarrow j$.

(ii) $\mathcal{G}$ is said to be undirected (UG) when all its edges are undirected.

Definitions 1.1.9 (Directed Edge and graph)**.** Let $\mathcal{G} = (\Gamma, E)$ be a graph,

(i) An edge (i, j) of $\mathcal{G}$ is said to be directed if $(i, j) \in E$ et $(j, i) \notin E$. It is then denoted by $i \rightarrow j$.

(ii) $\mathcal{G}$ is said to be directed (DG) when all its edges are directed.

Depending on whether it is directed or undirected, a graph is associated with a specific vocabulary. In an undirected graph, for any edge $i \leftrightarrow j$ in E, we will say that i and j are respective neighbors, and we will denote $\text{ne}_{\mathcal{G}}(i) = \{j \in \Gamma : (i, j) \in E\}$ the set of neighbors of i in $\mathcal{G}$. In the case of a directed graph, for any edge $i \rightarrow j$ in E, we will say that i is the parent of j (resp. j is the descendant of i), and we will denote $\text{pa}_{\mathcal{G}}(j) = \{i \in \Gamma : (i, j) \in E\}$ the set of parents of j in $\mathcal{G}$ (resp. $\text{de}_{\mathcal{G}}(i) = \{j \in \Gamma : (i, j) \in E\}$). However, it is possible to harmonize this information whatever the graph using the adjacency matrix defined by

$$A = (A_{ij})_{i,j \in \Gamma}, \qquad A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

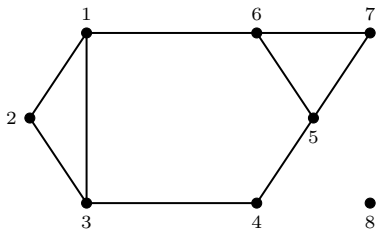
This one offers a simple representation of a graph that we will know oriented or not depending on whether its adjacency matrix is symmetric. The Figure 1.1 represents a directed graph and its undirected version with their adjacency matrices.

We will see later in the chapter that depending on the phenomenon being studied, it can be more convenient to manipulate undirected graphs than directed ones. It is therefore appropriate to introduce now on the the graph moralization criterion which enables to find the equivalent undirected form of an acyclic directed graph.

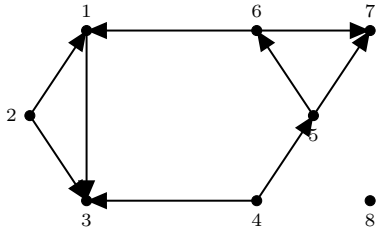
Proposition 1.1.10 (Directed acyclic graph)**.** *A directed graph $\mathcal{G}$ is said to be acyclic (DAG) only if $\mathcal{G}$ has a topological order. In other words, if it does not have a sequence of edges forming a loop in $\mathcal{G}$.*

Definition 1.1.11 (Moral graph)**.** The moral graph associated to an acyclic directed graph $\mathcal{G} = (\Gamma, E)$ is the undirected graph $\mathcal{G}^m = (\Gamma, E^m)$, such that the set of edges E^m is constructed as follow

$$\forall i, j \in \Gamma, \quad i \leftrightarrow j \quad \Longleftrightarrow \quad i \rightarrow j \text{ ou } j \rightarrow i \text{ ou } \{\exists k \in \Gamma : i \rightarrow k \text{ et } j \rightarrow k\}. \tag{8}$$



	1	2	3	4	5	6	7	8
1	0	1	1	0	0	1	0	0
2	1	0	1	0	0	0	0	0
3	1	1	0	1	0	0	0	0
4	0	0	1	0	1	0	0	0
5	0	0	0	1	0	1	1	0
6	1	0	0	0	1	0	1	0
7	0	0	0	0	1	1	0	0
8	0	0	0	0	0	0	0	0



	1	2	3	4	5	6	7	8
1	0	0	1	0	0	0	0	0
2	1	0	1	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	1	0	1	0	0	0
5	0	0	0	0	0	1	1	0
6	1	0	0	0	0	0	1	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0

FIGURE 1.1 : Examples of undirected (top) and directed (bottom) graphs with their adjacency matrices.

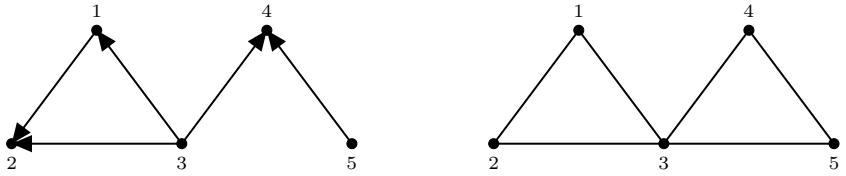


FIGURE 1.2 : Example of a directed graph with its associated moral graph.

1.2 Gaussian graphical models

In this section we will discuss the theoretical aspect of Gaussian graphical models. Let's take again our p-dimensional Gaussian random vector X, with zero mean, covariance $\Sigma \in \mathbb{S}^p_{++}$, and precision matrix $\Omega = \Sigma^{-1}$, and consider a graph $\mathcal{G} = (\Gamma, E)$ with $\Gamma = \{1, ..., p\}$. A Gaussian graphical model of X with respect to $\mathcal{G}$ is a probabilistic model, in which the graph $\mathcal{G}$ expresses the conditional dependence structure between the random variables of X. In particular, if $\mathcal{G}$ is undirected, it will contain the structure of correlations of the variables of X, whereas if it is directed it will contain the causal relations. We will see in this section that the definition of a graphical model relies on the respect of the pairwise Markov property, and we will study its interpretation in the framework of oriented and non-oriented graphs.

1.2.1 Markov properties for undirected graphs

Markov properties distinguish conditional independence structures represented by a graph, by exploiting the factorization 2 for different sets C on which the conditioning is based.

Definition 1.2.1. Let be a random vector X and an undirected graph $\mathcal{G}$. The vector X satisfies *the*

pairwise Markov property if for any pair of nodes i, j in Γ,

$$i \nleftrightarrow j \qquad \Longleftrightarrow \qquad X_i \perp\!\!\!\perp X_j | X_{\Gamma \setminus \{i,j\}}. \tag{9}$$

Definition 1.2.2. Let be a random vector X and an undirected graph $\mathcal{G}$. The vector X satisfies *the local Markov property* if for any nodes i in Γ,

$$X_i \perp\!\!\!\perp X_{\Gamma \setminus \mathrm{ne}_{\mathcal{G}}(i) \cup \{i\}} | X_{\mathrm{ne}_{\mathcal{G}}(i)}. \tag{10}$$

Definition 1.2.3. Let be a random vector X and an undirected graph $\mathcal{G}$. The vector X satisfies *the global Markov property* if for any subset of disjoint nodes $\{A, B, C\}$ in $\mathcal{G}$, such that in $\mathcal{G}$, A is separated from B given C (*i.e.* any path in $\mathcal{G}$ from a node in A to a node in B contains a node in C),

$$X_A \perp\!\!\!\perp X_B | X_C. \tag{11}$$

One can easily see that the global Markov property implies the local property which implies the pairwise property, and in well-defined cases, when the density of X is strictly positive, the reciprocal implications are also verified (*see observation 1.7.1 of* [MDLW18]). Thus, a Gaussian vector satisfying one of the Markov properties, satisfies them all.

1.2.2 Markov properties for directed acyclic graphs

Since the Markov properties arise from the factorization of the joint distribution over a set of conditional distributions, the presence of cycles in directed graphs setting poses a problem in this writing. Indeed, if we consider the loop $1 \to 2 \to 3 \to 1$, there is no coherent way to factorize the distribution of the vector of variables indexed by these nodes, since no node is guaranteed to d-separate its parents from its descendants. Therefore, we will restrict ourselves in this section to directed acyclic graphs. The factorization 2 is then slightly modified in the context of DAGs, in particular for any node i in Γ, the random variable X_i is associated to a conditional distribution with respect to the variables corresponding to the parents of i in the DAG. The joint distribution now verifies

$$f(X_1, ..., X_n) = \prod_{i \in \Gamma} f_{X_i | X_{pa(i)}}. \tag{12}$$

Definition 1.2.4. Let be a random vector X and a directed acyclic graph $\mathcal{G}$. The vector X satisfies *the pairwise Markov property* if for any pair of nodes i, j in Γ,

$$i \nrightarrow j \qquad \Longleftrightarrow \qquad X_i \perp\!\!\!\perp X_j | X_{\Gamma \setminus \mathrm{de}_{\mathcal{G}}(i) \cup \{i,j\}}. \tag{13}$$

Definition 1.2.5. Let be a random vector X and a directed acyclic graph $\mathcal{G}$. The vector X satisfies *the local Markov property* if for any node i in Γ,

$$X_i \perp\!\!\!\perp X_{\Gamma \setminus \mathrm{de}_{\mathcal{G}}(i) \cup \{i\}} | X_{\mathrm{pa}_{\mathcal{G}}(i)}. \tag{14}$$

Definition 1.2.6. Let be a random vector X and a directed acyclic graph $\mathcal{G}$. The vector X satisfies *the global Markov property* if for any subset of disjoint nodes $\{A, B, C\}$ in $\mathcal{G}$, such that in $\mathcal{G}$, A is d-separated from B given C (*i.e.* in the moral graph $\mathcal{G}^m$ associated with $\mathcal{G}$, A is separated from B given C),

$$X_A \perp\!\!\!\perp X_B | X_C. \tag{15}$$

The Markov properties associated with DAGs offer a more complex interpretation than those associated with UGs. In order to unify the two cases, we will exploit Theorem 3.5.2 of [Whi09] which states that a DAG satisfies the same Markov properties as its associated moral graph. Moreover, as we do not study the causal structures contained in DAGs, we will only be interested in the correlation structures contained in the associated moral graphs. Thus, in the rest of the manuscript we will focus on the case of undirected graphs.

1.2.3 Gaussian graphical models

Now that we have stated the properties of Markov, we can present the definition of a Gaussian graphical model.

Definition 1.2.7 (Gaussian graphical models)**.** Let be a Gaussian random vector X and an undirected graph $\mathcal{G}$. The distribution of X is a Gaussian graphical model with respect to $\mathcal{G}$ if it satisfies the pairwise Markov property.

It follows from this definition that the minimal graph $\mathcal{G}^*$ for which X satisfies the Markov properties describes the sparsity pattern of the precision matrix of X. In particular,

$$A_{ij} = A_{ji} = 0 \quad \Longleftrightarrow \quad \Omega_{ij} = \Omega_{ji} = 0, \tag{16}$$

where A is the adjacency matrix of $\mathcal{G}^*$. Thus, the estimation of the graph $\mathcal{G}^*$, and thus of the conditional independence structure of X, requires the estimation of the precision matrix Ω. There is a large variety of methods for estimating Ω, in this introduction we will discuss the maximum likelihood estimation and the Bayesian estimation.

1.3 Maximum likelihood estimation

Consider a sample of n independent observations random vectors according to our multivariate p-dimensional Gaussian distribution. Let us denote $X_i \in \mathbb{R}^p$ the values taken by individual i on the p random variables of X, and $\mathbf{X} = (X_i^t)_{1 \leqslant i \leqslant n}$ the matrix of observed data, of dimension $n \times p$. The

likelihood of associated to the data is given by

$$\begin{aligned}\mathcal{L}_n(\Omega) &= \prod_{i=1}^{n} f_X(X_i) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{\langle X_i, \Omega\, X_i\rangle}{2}\right) \\ &= \left[\frac{\det(\Omega)}{(2\pi)^p}\right]^{\frac{n}{2}} \exp\left(-\frac{\langle\!\langle \mathbf{X}^t\,\mathbf{X}, \Omega\rangle\!\rangle}{2}\right),\end{aligned} \tag{17}$$

with $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ refers to the Frobenius scalar product (*i.e.* $\langle\!\langle A, B\rangle\!\rangle = \mathrm{tr}(A^t\, B)$). We define $S \in \mathbb{R}^{p\times p}$, the empirical covariance matrix of the sample under the assumption of centered data, by

$$S = \frac{1}{n}\sum_{i=1}^{n} X_i\, X_i^t. \tag{18}$$

We can rewrite the likelihood with this quantity, which gives

$$\mathcal{L}_n(\Omega) = \left[\frac{\det(\Omega)}{(2\pi)^p}\right]^{\frac{n}{2}} \exp\left(-\frac{n}{2}\langle\!\langle S, \Omega\rangle\!\rangle\right). \tag{19}$$

The estimator of Ω is obtained by maximizing $L_n(\Omega)$. In order to simplify the problem in an additive form, we will instead use the equivalent optimization problem which consists in minimizing the negative log-likelihood. By deleting from (19) the terms not depending on Ω, up to a coefficient we obtain

$$\ell\ell_n(\Omega) = -\ln\det(\Omega) + \langle\!\langle S, \Omega\rangle\!\rangle. \tag{20}$$

Thus, the optimization problem boils down to

$$\hat{\Omega}_{\mathrm{or}} = \arg\min_{\Omega\in\Theta_{\mathrm{or}}} \ell\ell_n(\Omega). \tag{21}$$

Moreover, let us note that in the Gaussian graphical model setting associated with a graph $\mathcal{G} = (\Gamma, E)$,

$$\Theta_{\mathrm{or}} = \{\Omega \in \mathbb{S}^p_{++} \,|\, \Omega_{ij} = 0,\ \ \forall\, i \neq j \ \text{ with } \ (i,j) \notin E\}.$$

Theorem 1.3.1. *In a saturated Gaussian graphical model (i.e. all the nodes of the graph are connected two by two), the maximum likelihood estimator exists only if $n \geqslant p$. In this case it is given by*

$$\hat{\Omega} = S^{-1}. \tag{22}$$

Proof. **Convexity of** (20)**:** Proposition 9.2.1 of [MDLW18] states that the function $\ln\det(Y) - \langle\!\langle S, Y\rangle\!\rangle$ is concave on $\mathbb{S}^p_{++}$ (see the associated proof). This implies that the negative log-likelihood is convex on $\mathbb{S}^p_{++}$. Since the set Θ_{or}, representing the set of possible precision matrices of the model, is a convex cone in $\mathbb{S}^p_{++}$, then the problem (21) is convex on Θ_{or}.

Estimator:

$$\begin{aligned} & \frac{\partial}{\delta\Omega}\ell\ell_n(\Omega) = 0 \\ \iff \quad & -\frac{\delta}{\delta\Omega}\ln\det(\Omega) + \frac{\delta}{\delta\Omega}\langle\!\langle S, \Omega\rangle\!\rangle = 0 \\ \iff \quad & -\Sigma + S = 0 \\ \iff \quad & S = \Sigma. \end{aligned}$$

Condition on the sample: S^{-1} can only be a solution of the optimization problem if it is in $\mathbb{S}^p_{++}$, yet S is singular when $n < p$. □

Theorem 1.3.1 raises two problems. First, the maximum likelihood estimator does not exist in high-dimensional setting, but also the latter is not sparse in the classical case. This method of estimation does not meet the needs generally encountered in practice, in particular the treatment of data with a number of predictors largely superior to the number of observations, and for which we are looking for a sparse solution.

1.4 Penalized Gaussian graphical models

The respect of the criterion of parsimony is a major concern when creating a high dimensional model. We are looking for a relevant solution involving a minimum of parameters. Many studies dealing with this issue have been done in the context of Gaussian graphical models. The idea is to select a sparse undirected graph, which results in the estimation of a sparse precision matrix, its zeros expressing conditional independences between the variables. We can cite for example the traditional approach, greedy stepwise forward-backward selection, which consists in the iterative selection and/or deletion of the neighbors of each node using multiple tests (*e.g.* [Dem72, Wer76, Edw00]). Another approach is to represent each variable as a linear combination of the other variables of X, and then to perform a penalized regression (a Lasso-type for example) in order to recover the set of neighbors of each node of the graph (*e.g.* [MB06, CLL11]).
In this section, we will focus on the penalized maximum likelihood approach. The idea is to take the optimization problem (21) to which we add a penalty on the precision matrix. This means finding $\Omega \in \mathbb{S}^p_{++}$ which minimizes the convex objective function

$$\ell\ell_n(\Omega) = -\ln\det(\Omega) + \langle\!\langle S, \Omega\rangle\!\rangle + \lambda\,\mathrm{pen}(\Omega). \tag{23}$$

The function $\mathrm{pen}(\Omega)$ is the penalty function applied to the precision matrix. When one wishes to induce sparsity within it, $|\Omega|_1$ the sum of the absolute value of each element of Ω, or $|\Omega|_1^-$ this sum deprived of the diagonal elements, are well-known as natural choices. However, note that if some authors use the penalty $|\Omega|_1$, the estimation algorithms they present are built so as not to penalize the diagonal elements. The parameter λ enables to give more or less weight to the penalty function, thus varying the degree of sparsity and shrinkage within the precision matrix. When $\lambda = 0$ we obtain the maximum likelihood estimator. The practitioner can choose the value of this parameter in order to obtain an estimator $\hat{\Omega} \in \mathbb{S}^p_{++}$, but also

according to the desired degree of sparsity, or the one considered optimal according to a model selection criterion (*e.g.* AIC, BIC, cross-validation error).
The idea of using the ℓ_1 regularization as a penalty function is clearly inspired by Tibshirani's Lasso tibshirani96, in order to take advantage of the sparsity and shrinkage qualities related to this estimator. Efficient algorithms exist to find solutions to the optimization problem (23). Let us cite some of them.

The block coordinate descent of Banerjee *et al.* [BEGD08]. In their paper, Banerjee *et al.* consider the penalty $|\Omega|_1$, and treat the problem (23) via its dual form below, which estimates the covariance matrix instead of its inverse.

$$\hat{\Sigma} = \arg\max_{W} \{\ln\det(W) : |W - S|_\infty \leqslant \lambda\}, \tag{24}$$

with $W = S + U$, and U a symmetric matrix. They present a block descent algorithm for the problem (24) which consists in a recursive update of the maximum likelihood estimator. At each iteration, they consider the matrix $W^{(j-1)}$, which is the estimator updated in the previous step, and the empirical covariance matrix S which verify the following decompositions:

$$W^{(j-1)} = \begin{pmatrix} W_{\setminus j} & W_j^- \\ (W_j^-)^t & W_{jj} \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} S_{\setminus j} & S_j^- \\ (S_j^-)^t & S_{jj} \end{pmatrix}$$

where $A_{\setminus j}$ corresponds to the matrix A deprived of the row and the column j, and A_j^- corresponds to the column j of A deprived of the diagonal element A_{jj}. In order to update $W^{(j)}$, they then estimate W_j^- by solving the following quadratic problem:

$$W_j^- = \arg\min_{y} \{y^t\,(W_{\setminus j}^{(j-1)})^{-1}\,y : \|y - S_j^-\|_\infty \leqslant \lambda\}. \tag{25}$$

This estimation method echoes that of Meinhausen and Bühlmann [MB06], however the authors guarantee here the uniqueness of the solution obtained at each regression, and the estimator is updated iteratively until convergence. This algorithm offers a computational complexity $\mathcal{O}(Kp^4)$ where K denotes the number of iterations before convergence.

Nesterov's first order method by Banerjee *et al.* [BEGD08]. In the same paper, Banerjee *et al.* also presented an estimation algorithm based on the Nesterov's first order method. The goal is not really to provide another algorithm, but to go through Nesterov's smooth minimization method in order to obtain a more rigorous and accurate complexity estimation than the block descent one. First, the problem (23) had to be reformulated as a non-smooth optimization problem adapted to the Nesterov method. For this purpose, the authors have shown that when we impose a bound on the eigenvalues of Ω, the problem (23) can be expressed by

$$\min_{\Omega \in \mathcal{Q}_1} -\ln\det(\Omega) + \langle\!\langle S, \Omega \rangle\!\rangle + \max_{u \in \mathcal{Q}_2} \langle \Omega, u \rangle_2\,, \tag{26}$$

avec $\mathcal{Q}_1 = \{\Omega : aI \preceq \Omega \preceq bI\}$, and $\mathcal{Q}_2 = \{u : \|u\|_\infty \leqslant \lambda\}$. We will note f the function to minimize in (26). Once the reference format of the Nesterov problem is obtained, the first order method is articulated in two steps. First, a smooth approximation of the objective function f is constructed by adding a convex penalty on u ; then the first order optimization algorithm, presented in [Yur83], is applied to this new function. This algorithm offers a complexity $\mathcal{O}(p^{4.5}/\epsilon)$ where $\epsilon > 0$ denotes the desired accuracy on the optimization problem (23).

Block coordinate descent by Friedman *et al.* [FHT08]. In their paper, Friedman *et al.* consider the $|\Omega|_1^-$ penalty and present a blockwise coordinate descent algorithm that relies on the same principle as that of Banerjee *et al.*, but with some variants that make it computationally favorable. In particular, they exploit a point raised by the first authors, namely that solving the problem (25) is equivalent to solving the dual problem

$$W_j^- = \min_x |Q\,x - b|_2^2 + \lambda |x|_1, \tag{27}$$

with Q the square root of $W_{\setminus j}$ and $b = \frac{1}{2}\,Q^{-1}\,S_j^-$. The estimation algorithm becomes mainly a recursive Lasso, where at each step the diagonal element is omitted from the regression problem to prevent it from being penalized. The computational complexity of this algorithm is $\mathcal{O}(p^3)$ under a dense model (*i.e.* the covariance matrix is full), and is less in the sparse case. To complete this non-exhaustive list we can also mention the neighborhood selection of Meinshausen and Bühlmann [MB06], which consists in searching for the neighbors of each variable by performing a Lasso regression expressing the latter as a function of the other variables ; or the smooth minimization approach of Lu [Lu09], also based on Nesterov's first order method, but offering a more attractive complexity than the algorithm of the authors of [BEGD08]. In addition to the properties of these algorithms reported in each article, we can find the theoretical guarantees associated with this type of penalization in the paper of Ravikumar *et al.* [RWRY11].

1.5 Partial Gaussian graphical model

In this section we will present the partial Gaussian graphical model (PGGM), introduced by Sohn and Kim [SK12], and Yuan and Zhang [YZ14] . We will first formulate the motivation of this model by highlighting the link between linear regression and GGM, then we will discuss its theoretical and inferential aspects.

1.5.1 Link with linear regression

We are now in the linear regression setting where we want to express multiple responses $Y \in \mathbb{R}^q$ from a set of predictors $X \in \mathbb{R}^p$. More precisely we are looking for the matrix of regression coefficients $B \in \mathbb{R}^{p\times q}$ such that

$$Y = B^t\,X + E, \tag{28}$$

with $E \sim \mathcal{N}_q(0, R)$ a Gaussian noise. Let $Z = (Y, X) \sim \mathcal{N}_{q+p}(0, \Sigma)$ be the Gaussian vector associated to the joint distribution of Y and X. The distribution of Z is a Gaussian graphical model for which the associated graph is determined according to the conditional independence structure provided by the

precision matrix $\Omega = \Sigma^{-1}$. We observe once again the following block decompositions

$$Z = \begin{pmatrix} Y \\ X \end{pmatrix}, \quad \Omega = \begin{pmatrix} \Omega_y & \Delta \\ \Delta^t & \Omega_x \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_y & \Sigma_{yx} \\ \Sigma_{yx}^t & \Sigma_x \end{pmatrix}$$

where $\Omega_y \in \mathbb{S}_{++}^q$, $\Delta \in \mathbb{R}^{q \times p}$ and $\Omega_x \in \mathbb{S}_{++}^p$, and where the same goes for Σ_x, Σ_{yx} and Σ_x. Moreover by Schur, the precision matrix Ω satisfies, by blockwise inversion,

$$\Omega_y^{-1} = \Sigma_y - \Sigma_{yx}\,\Sigma_x^{-1}\,\Sigma_{yx}^t \quad \text{and} \quad \Delta = -(\Sigma_y - \Sigma_{yx}\,\Sigma_x^{-1}\,\Sigma_{yx}^t)^{-1}\,\Sigma_{yx}\,\Sigma_x^{-1}. \tag{29}$$

If for an individual i we observe the values X_i on the p predictors of X, the proposition 1.1.5 notifies that the conditional distribution of Y_i given X_i satisfies

$$Y_i \,|\, X_i \;\sim\; \mathcal{N}(-\Omega_y^{-1}\,\Delta\,X_i,\, \Omega_y^{-1}). \tag{30}$$

By associating the conditional distribution (30) with the linear writing (28), we obtain

$$B = -\Delta^t\,\Omega_y^{-1}, \qquad R = \Omega_y^{-1}. \tag{31}$$

The reparametrization (31) gives a new light on the multiple-output regression. Whereas B contains direct and indirect links between the predictors and the responses (due *e.g.* to strong correlations among the variables), Δ only contains direct links, as it is shown by the graphical models theory. In other words, the direct links are closely related to the concept of partial correlations between X and Y. For example, the direct link between predictor k and response ℓ may be evaluated through the partial correlation $\mathbb{C}\mathrm{orr}(Y_\ell, X_k \,|\, Y_{\neq \ell}, X_{\neq k})$ contained, apart from a multiplicative coefficient, in the ℓ-th row and k-th column of Δ (see section 1.1.2), with the particularly interesting consequence that the support of Δ is sufficient to identify direct relationships between X and Y. Thus, when Y is multidimensional, it makes more sense to study the conditional independence between predictors and responses through Δ, B being influenced by the indirect links that propagate through Ω_y.

In this vein, Sohn and Kim [SK12] and Yuan and Zhang [YZ14] proposed a partial estimation of the Gaussian graphical model (PGGM). The objective reduces to the estimation of the direct links Δ together and the conditional precision matrix of the responses Ω_y, possibly penalized. This approach has a huge advantage over GGMs from an inferential point of view. Since we are not interested in partial correlations between the predictors, we do not need to impose any assumption on the structure of X, notably its degree of sparsity. Moreover, since the size of X is, in general, much larger than that of Y (*e.g.* omics data processing, imaging, NLP), the estimation of the full precision matrix Ω is highly impacted by the estimation of Ω_x. Therefore, the bias induced in the estimation of Ω_x, due in particular to the difficulties related to the high dimensional analysis, will affect the estimates of Δ and Ω_y. Finally, it is obvious that the computational complexity, which we have not discussed so far since it is closely related to the estimation algorithms, will be strongly reduced with a partial estimation.

1.5.2 Penalized partial likelihood

Let's consider again our n observations. This time for each individual we observe the pair (Y_i, X_i) of values taken on the predictors and the responses. The empirical covariance S satisfies the following block decomposition

$$S = \begin{pmatrix} S_y & S_{yx} \\ S_{yx}^t & S_x \end{pmatrix}$$

with $\quad S_y = \frac{1}{n}\sum_{i=1}^{n} Y_i\, Y_i^t, \quad S_{yx} = \frac{1}{n}\sum_{i=1}^{n} Y_i\, X_i^t \quad$ and $\quad S_x = \frac{1}{n}\sum_{i=1}^{n} X_i\, X_i^t$.

In the following Proposition, L denotes the negative of the logarithm of the likelihood function corresponding to the GGM.

Proposition 1.5.1 (Proposition 1 of [YZ14]). *Under the transformation* $\tilde{\Omega}_x = \Omega_x - \Delta^t\, \Omega_y^{-1}\, \Delta$ *we have,*

$$L(\Omega_y, \Delta, \Omega_x) \;=\; \tilde{L}(\Omega_y, \Delta, \tilde{\Omega}_x) = L_{pa}(\Omega_y, \Delta) + H(\tilde{\Omega}_x), \tag{32}$$

where $H(\tilde{\Omega}_x) = -\ln\det(\tilde{\Omega}_x) + \langle\!\langle S_x, \tilde{\Omega}_x \rangle\!\rangle$ *and*

$$L_{pa}(\Omega_y, \Delta) \;=\; -\ln\det(\Omega_y) + \langle\!\langle S_y, \Omega_y \rangle\!\rangle + 2\langle\!\langle S_{yx}, \Delta \rangle\!\rangle + \langle\!\langle S_x, \Delta^t\, \Omega_y^{-1}\, \Delta \rangle\!\rangle \tag{33}$$

Moreover, $L_{pa}(\Omega_y, \Delta)$ *is convex.*

Proof. Refer to the proof of the proposition provided in Appendix A.1 of [YZ14]. □

The proposition 1.5.1 enables to decompose the optimization problem (21) into two distinct problems by the reformulation (32). Since $\tilde{L}$ is jointly convex, $\hat{\Omega}$ is obtained on the one hand by finding the estimators of Ω_y and Δ by minimizing the partial likelihood L_{pa}, and on the other hand by finding the estimator of $\tilde{\Omega}_x$, from which we will be able to deduce the estimator of Ω_x, by minimizing the function H. After having taken care to reintegrate the penalty functions, we obtain the new convex objective function of the PGGM,

$$\begin{aligned} L_{pa}(\Omega_y, \Delta) \;=\; & -\ln\det(\Omega_y) + \langle\!\langle S_y, \Omega_y \rangle\!\rangle + 2\,\langle\!\langle S_{yx}, \Delta \rangle\!\rangle \\ & +\; \langle\!\langle S_x, \Delta^t\, \Omega_y^{-1}\, \Delta \rangle\!\rangle + \lambda\, \mathrm{pen}(\Omega_y) + \mu\, \mathrm{pen}(\Delta) \end{aligned} \tag{34}$$

defined over $(\Omega_y, \Delta) \in \Theta = \mathbb{S}_{++}^q \times \mathbb{R}^{q\times p}$ for some usual penalty functions. It is worth noting that $\mathrm{pen}(\Delta)$ often plays a crucial role in modern statistics dealing with high-dimensional predictors, and the natural choice is $|\Delta|_1$ to get sparsity. However, since the number of responses is usually small, we can choose $\lambda = 0$ and impose no sparsity structure within Ω_y. In the seminal papers [SK12] and [YZ14], the authors consider $|\Omega_y|_1$ and $|\Omega_y|_1^-$ for $\mathrm{pen}(\Omega_y)$, respectively. Consider $\theta = (\Omega_y, \Delta) \in \Theta$ the matrix of dimension $(q\times(q+p))$ of the model parameters. The estimator $\widehat{\theta}$ of θ is obtained by solving the following optimization problem

$$\widehat{\theta} \;=\; (\widehat{\Omega}_y, \widehat{\Delta}) \;=\; \underset{\theta\in\Theta}{\arg\min}\; L_{pa}(\Omega_y, \Delta). \tag{35}$$

Yuan and Zhang propose in [YZ14] an algorithm for estimating the problem (35) with complexity

$\mathcal{O}(p^3+p^2q+pq\min\{n,q\})$. It is competitive against its counterparts in the GGM setting, the best having a complexity $\mathcal{O}((p+q)^3)$ (see section 1.4). This algorithm adopts a block coordinate descent approach, that exploits the joint convexity property of (35) by alternating between solving the following two subproblems

$$\Omega_y^{(t+1)} = \underset{\Omega_y \in \mathbb{S}_{++}^q}{\arg\min}\ L_{pa}(\Omega_y, \Delta^{(t)}) \quad et \quad \Delta^{(t+1)} = \underset{\Delta \in \mathbb{R}^{q\times p}}{\arg\min}\ L_{pa}(\Omega_y^{(t+1)}, \Delta). \tag{36}$$

Yuan and Zhang have also provided in [YZ14] theoretical guarantees on the penalized PGGM estimator, which takes the form of an upper bound on the estimation error. More precisely, they show that when n is large and $p \gg n$, with high probability, the estimation error is given by

$$\|\widehat{\theta} - \theta^*\|_F \lesssim \sqrt{\frac{|S|\ \ln p}{n}}, \tag{37}$$

where θ^* are the true parameters of the model, and $|S|$ is the cardinal of the support of θ^*. We will re-demonstrate this result in Chapter 2 when we will provide the model with an additional structural regularization.

Moreover, while Bayesian GGM has been extensively studied in the literature (see e.g. Maathuis et al., 2018, Chap. 10), the Bayesian counterpart of PGGM, in contrast, to the best of our knowledge, the Bayesian counterpart of PGGM has not yet been developed. This approach will be the subject of our Chapter 3.

Chapter 2

A PARTIAL GRAPHICAL MODEL WITH A STRUCTURAL PRIOR ON THE DIRECT LINKS

In the tree of high-dimensional regression problems, many of them embed a correlation structure within the predictors. This is notably the case in genomics with gene regulatory networks that provide information about interactions between genes and their joint actions within cells; similarly, in signal theory predictors often represent a continuous phenomenon so that consecutive variates act together, this can be observed for example with adjacent pixels in an image. Well-known methods in the literature, such as the Group Lasso of Yuan and Lin [YL06], the blockwise sparse regression of Kim *et al.* [KKK06], or the sparse-group Lasso of Simon *et al.* [SFHT13], integrate in their estimation process a structuring in groups of variables supposedly correlated. One of the main limitations of these methods is their lack of flexibility in group construction. Indeed, each variable is assigned to one group, assuming that it has no significant correlation with variables in other groups. To get around this assumption, which in practice is not always verified (*e.g.* image processing, genetic data, etc.), studies have been carried out to extend this problem by allowing overlapping groups. To that extent we can cite the CAP penalties of Zhao *et al.* [ZRY09], the Graph Lasso and Group Lasso with overlap of Jacob *et al.* [JOV09], the Structured-Lasso and Intersected Structured-Lasso of Jenatton *et al.* [JAB11], or the Tree-Guided Group Lasso by Kim and Xing [KX12]. Although some of these methods borrow elements from graph theory to define complex groupings of predictors, the interactions between the responses remain unexplored. In this vein, graphical models naturally appear as a way to explore, and more precisely Gaussian graphical models with a structural penalty. This is precisely the subject of the article [CMHR17] by Chiquet *et al.*. Inspired by the structural penalizations proposed by Slawski [SZCT10] and Li and Li [LL10] in their structured versions of the Elastic Net estimator, and exploiting the partial reparametrization of the Gaussian graphical models of Sohn and Kim [SK12] and Yuan and Zhang [YZ14], they propose the Spring method, which consists of a PGGM with an additional structural penalty on direct links.

In this chapter we present the results of a collaboration with Frédéric Proïa and Pascal Jézéquel, which has been published in ESAIM Probability and Statistics in 2021 [OOJP21][1]. Our work is articulated on two axes. The first axis consists in the development of an estimation procedure for a penalized PGGM, named GenGm, which is based on a combination of the approaches of Yuan and Zhang [YZ14] and Chiquet *et al* . [CMHR17]. We bring two types of penalty to our optimization problem: a standard ℓ_1

1. The codes and the dataset are available at `https://github.com/EuniceOkome/StructPGGM`

penalty on Δ and Ω_y to induce sparsity within the partial correlations, and possibly within the responses precision matrix; and a structural penalty reflecting a generalized Gaussian prior on the direct links, which guides the sparsity pattern from a prior knowledge of the interactions between predictors. The second axis aims at providing a theoretical guarantee for our method, taking the form of an upper bound on the estimation error arising with high probability, provided that the model is suitably regularized.

This chapter is organized in four parts. In Section 2.1 we start with a reminder of the mathematical formalization of PGGM proposed by [YZ14]. In Section 2.2, we introduce the model, consisting in putting a generalized Gaussian prior on the direct links before the procedure of estimation of Ω_y and Δ, and we detail the new objective function. Then, in Section 2.3 we provide error bounds for our estimates and prove our results in Section 2.4. Section 2.5 is devoted to empirical considerations. We explain how we deal with the minimization of the new objective and we test the method on simulations first, and next on a real dataset (a Canadian average annual weather cycle, see *e.g.* [RS06]). Finally, we close the chapter with a short conclusion in the section 2.6.

2.1 Introduction

Let us consider, now and in all the study, the sample of n independent observations (Y_i, X_i), and the empirical covariances denoted by

$$S_y = \frac{1}{n}\sum_{i=1}^{n} Y_i\, Y_i^t, \qquad S_{yx} = \frac{1}{n}\sum_{i=1}^{n} Y_i\, X_i^t \qquad \text{et} \qquad S_x = \frac{1}{n}\sum_{i=1}^{n} X_i\, X_i^t. \tag{1}$$

We are in the PGGM setting discussed in Section 1.5. Without going into the details already covered, let's take up the objective function (34), where we consider the penalties

$$\mathrm{pen}(\Omega_y) = |\Omega_y|_1^- \qquad \text{and} \qquad \mathrm{pen}(\Delta) = |\Delta|_1 \tag{2}$$

which correspond to the PGGM (Gm) of [YZ14]. The Spring (Spr) of [CMHR17] can also be seen as a PGGM but with no penalty on Ω_y (replaced with an additional structuring one on Δ, we will come back to this point thereafter), so for (Spr) we may consider $\lambda = 0$. The generalized procedure (GenGm) at the heart of the study relies on a combination between these two approaches. We will see in due time that we keep both the penalties of (Gm) and the structuring one of (Spr) on Δ.

2.2 A generalized Gaussian prior on the direct links

As explained in the introduction, our method integrates a structural regularization on the direct links Δ through a generalized Gaussian prior. This section will be devoted to the presentation of this prior and the functioning of the associated structural regularization.

Remark 2.2.1. Let us specify that although we use the term *prior* here, we do not follow a Bayesian approach, the estimation step not being based on the search for an posterior distribution. We will carry out a more detailed study of this approach in Chapter 3.

2.2.1 Generalized Gaussian distribution

Recall that the density of a multivariate Gaussian distribution $\mathcal{N}(0, V)$, with zero mean[2] and covariance $V \in \mathbb{S}_{++}^d$, is given by

$$\forall\, z \in \mathbb{R}^d, \qquad f_V(z) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(V)}} \exp\left(-\frac{\langle z, V^{-1}z\rangle}{2}\right).$$

The generalization of this law involves two new parameters. A scale parameter $m \in]0, \infty[$, and a shape parameter $\beta \in]0, \infty[$. However, in our study we will only consider the cases where $m = 1$. According to the definition given in formulas (1)-(2) of [PBTB13], the density of a d-dimensional multivariate generalized Gaussian distribution $\mathcal{GN}(0, 1, V, \beta)$ takes the form of

$$\forall\, z \in \mathbb{R}^d, \qquad f_{V,\beta}(z) = \frac{\beta\,\Gamma(\frac{d}{2})}{\pi^{\frac{d}{2}}\,\Gamma(\frac{d}{2\beta})\,2^{\frac{d}{2\beta}}\sqrt{\det(V)}} \exp\left(-\frac{\langle z, V^{-1}z\rangle^{\beta}}{2}\right)$$

where Γ is the Euler Gamma function.
We clearly recognize the Gaussian $\mathcal{N}(0, V)$ setting for $\beta = 1$. Moreover, for $\beta = 1/2$, it can be seen as a multivariate Laplace distribution whereas it is known to converge to some uniform distribution as $\beta \to +\infty$. The marginal shapes ($d = 1$ and $V = 1$) of the distribution are represented on Figure 2.1, depending on whether $\beta < 1$, $\beta = 1$ or $\beta > 1$.

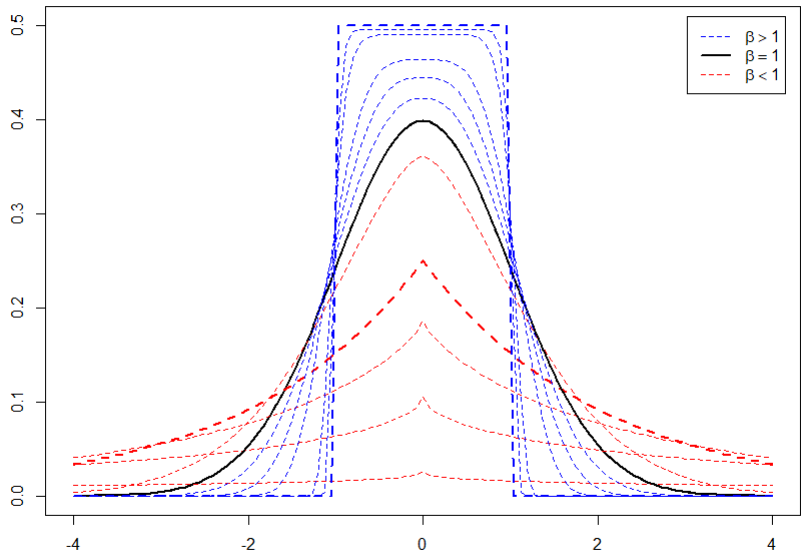


FIGURE 2.1 : Marginal shape of the generalized Gaussian distribution ($d = 1$ and $V = 1$) for some $\beta < 1$ (dotted red), $\beta = 1$ (black) and some $\beta > 1$ (dotted blue). The noteworthy cases $\beta = 1/2$ (Laplace), $\beta = 1$ (Gaussian) and $\beta = +\infty$ (uniform) are highlighted.

2. We consider here the centered case because the data are assumed to be normalized.

2.2.2 Formulation of the penalty

Our approach consists in integrating information on the structure of the interactions between the predictors within the optimization problem. To this end, although we are not in the Bayesian framework, we borrow a convention from this inference in order to formalize our structural penalization. In particular, the usual Bayesian approach for multiple-output Gaussian regression having B as matrix of coefficients and R as noise variance consists in a conjugate prior $\mathrm{vec}(B) \sim \mathcal{N}(b, R \otimes L^{-1})$ for some information matrix $L \in \mathbb{S}^p_{++}$ and a centering value b (see *e.g.* Section 2.8.5 of [RAM12]). In the PGGM reformulation, we have $R = \Omega_y^{-1}$ and $B = -\Delta^t \Omega_y^{-1}$ as explained in Section 1.5, and of course we shall choose $b = 0$ to meet our purposes. Thus, if we pose $L \in \mathbb{S}^p_{++}$ the structural information matrix

$$\mathrm{vec}(\Delta^t) = -(\Omega_y \otimes I_p)\,\mathrm{vec}(B) \;\sim\; \mathcal{N}(0, \Omega_y \otimes L^{-1})$$

is a natural prior for the direct links (this is in particular the choice of the authors of [CMHR17]). Indeed,

$$\begin{aligned}\mathrm{vec}(B) &= \mathrm{vec}(-\Delta^t \Omega_y^{-1}) \\ &= -(\Omega_y^{-1} \otimes I_p)\,\mathrm{vec}(\Delta^t),\end{aligned}$$

from which we get,

$$\begin{aligned}\mathrm{vec}(\Delta^t) &= -(\Omega_y^{-1} \otimes I_p)^{-1}\,\mathrm{vec}(B) \\ &= -(\Omega_y \otimes I_p)\,\mathrm{vec}(B) \\ &\sim \mathcal{N}(-(\Omega_y \otimes I_p)\,b, (\Omega_y \otimes I_p)\,(R \otimes L^{-1})(\Omega_y \otimes I_p)^t) \\ &\sim \mathcal{N}(-(\Omega_y \otimes I_p)\,b, \Omega_y\,R\,\Omega_y \otimes I_p\,L^{-1}\,I_p).\end{aligned}$$

Following the same logic, let us choose $\Omega_y \otimes L^{-1}$ for scatter parameter and suppose that

$$\mathrm{vec}(\Delta^t) \;\sim\; \mathcal{GN}(0, 1, \Omega_y \otimes L^{-1}, \beta). \tag{3}$$

In this way, we can play on the intensity of the constraint we want to bring on Δ, from a non-informative prior ($\beta = \infty$) to quasi-boundedness ($\beta \to 0$) through Laplace ($\beta = 1/2$) and Gaussian distributions ($\beta = 1$). This prior entails an additional smooth term acting as a structural penalization in the objective (34) that becomes

$$\begin{aligned}L_{pa}(\Omega_y, \Delta) \;=\; &-\ln\det(\Omega_y) + \langle\!\langle S_y, \Omega_y \rangle\!\rangle + 2\,\langle\!\langle S_{yx}, \Delta \rangle\!\rangle \\ &+ \langle\!\langle S_x, \Delta^t \Omega_y^{-1}\,\Delta \rangle\!\rangle + \eta\,\langle\!\langle L, \Delta^t \Omega_y^{-1}\,\Delta \rangle\!\rangle^{\beta} + \lambda\,|\Omega_y|_1^- + \mu\,|\Delta|_1,\end{aligned} \tag{4}$$

with three regularization parameters (λ, μ, η). The smooth penalization lends weight to the prior on Δ and thereby plays on the extent of shrinkage and structuring through β, whereas $|\Delta|_1$ and $|\Omega_y|_1^-$ are designed to induce sparsity. One can note that this is closely related to the log-likelihood of a hierarchical model of the form

$$\begin{cases} Y_i \,|\, X_i, \Delta \;\sim\; \mathcal{N}(-\Omega_y^{-1}\,\Delta\, X_i,\, \Omega_y^{-1}) \\ \mathrm{vec}(\Delta^{\,t}) \;\sim\; \mathcal{GN}(0, 1, \Omega_y \otimes L^{-1}, \beta) \end{cases}$$

where the emphasis is on Δ in the prior and Ω_y remains a fixed parameter. We can indeed see that

$$f(Y_i, \Delta | X_i, \Omega_y) = f(Y_i | X_i, \Omega_y, \Delta)\, f(\Delta | \Omega_y).$$

We know that $Y_i \,|\, X_i, \Omega_y, \Delta \;\sim\; \mathcal{N}(-\Omega_y^{-1}\,\Delta\, X_i,\, \Omega_y^{-1})$. Thus,

$$\begin{aligned} f(Y_i | X_i, \Omega_y, \Delta) &= \frac{1}{(2\pi)^{\frac{p}{2}}\, \det(\Omega_y^{-1})^{\frac{1}{2}}} \exp\left(- \frac{\big\langle (Y_i + \Omega_y^{-1}\Delta X_i),\, \Omega_y (Y_i + \Omega_y^{-1}\Delta X_i) \big\rangle}{2} \right) \\ &\propto \det(\Omega_y)^{\frac{1}{2}} \exp\left(- \frac{\big\langle (Y_i + \Omega_y^{-1}\Delta X_i),\, \Omega_y (Y_i + \Omega_y^{-1}\Delta X_i) \big\rangle}{2} \right). \end{aligned}$$

We also know that $\mathrm{vec}(\Delta^{\,t}) \;\sim\; \mathcal{GN}(0, 1, \Omega_y \otimes L^{-1}, \beta)$. Thus,

$$\begin{aligned} f(\Delta | \Omega_y) &\propto \frac{1}{\det(\Omega_y \otimes L^{-1})^{\frac{1}{2}}} \exp\left(- \frac{\big\langle \mathrm{vec}(\Delta^{\,t}), (\Omega_y^{-1} \otimes L)\, \mathrm{vec}(\Delta^{\,t}) \big\rangle^{\beta}}{2} \right) \\ &\propto \det(\Omega_y)^{-\frac{p}{2}}\, \det(L)^{\frac{q}{2}} \exp\left(- \frac{\big(\mathrm{vec}(\Delta^{\,t})^t\, \mathrm{vec}(L\, \Delta^{\,t}\, \Omega_y^{-1}) \big)^{\beta}}{2} \right) \\ &\propto \det(\Omega_y)^{-\frac{p}{2}} \exp\left(- \frac{\big\langle\!\big\langle L,\, \Delta^{\,t}\, \Omega_y^{-1} \Delta \big\rangle\!\big\rangle^{\beta}}{2} \right). \end{aligned}$$

We recognize the structural term that appears in our penalty. Moreover,

$$f(Y_i, \Delta | X_i, \Omega_y) \propto \det(\Omega_y)^{\frac{1}{2}} \exp\left(- \frac{\big\langle (Y_i + \Omega_y^{-1}\Delta X_i),\, \Omega_y (Y_i + \Omega_y^{-1}\Delta X_i) \big\rangle + \big\langle\!\big\langle L,\, \Delta^{\,t}\, \Omega_y^{-1} \Delta \big\rangle\!\big\rangle^{\beta}}{2} \right),$$

from which we get

$$\begin{aligned} \mathcal{L}_n(\Omega_y, \Delta) &= \prod_{i=1}^{n} f(Y_i, \Delta | X_i, \Omega_y) \\ &\propto \det(\Omega_y)^{\frac{n}{2}} \exp\left(- \frac{\sum_{i=1}^{n} \big\langle (Y_i + \Omega_y^{-1}\Delta X_i),\, \Omega_y (Y_i + \Omega_y^{-1}\Delta X_i) \big\rangle}{2} \right) \\ &\quad \times \exp\left(- \frac{n \big\langle\!\big\langle L,\, \Delta^{\,t}\, \Omega_y^{-1} \Delta \big\rangle\!\big\rangle^{\beta}}{2} \right) \\ &\propto \det(\Omega_y)^{\frac{n}{2}} \exp\left(- \frac{\big\langle\!\big\langle (Y + \Omega_y^{-1}\Delta X),\, \Omega_y\, (Y + \Omega_y^{-1}\Delta X) \big\rangle\!\big\rangle}{2} \right) \\ &\quad \times \exp\left(- \frac{n \big\langle\!\big\langle L,\, \Delta^{\,t}\, \Omega_y^{-1} \Delta \big\rangle\!\big\rangle^{\beta}}{2} \right), \end{aligned}$$

and so

$$
\begin{aligned}
\ell\ell_n(\Omega_y, \Delta) = \; & \text{cst} + \frac{n}{2} \ln\det(\Omega_y) - \frac{1}{2} \langle\!\langle (Y + \Omega_y^{-1}\Delta X),\, \Omega_y\,(Y + \Omega_y^{-1}\Delta X) \rangle\!\rangle \\
& - \frac{n}{2} \langle\!\langle L,\, \Delta^{\,t}\, \Omega_y^{-1}\, \Delta \rangle\!\rangle^{\beta} \\
= \; & \text{cst} + \frac{n}{2} \ln\det(\Omega_y) - \frac{1}{2} \langle\!\langle Y\, Y^{\,t},\, \Omega_y \rangle\!\rangle - \frac{1}{2} \langle\!\langle \Omega_y^{-1}\, \Omega_y\, Y\, X^{\,t},\, \Delta \rangle\!\rangle - \frac{1}{2} \langle\!\langle X\, Y^{\,t}\, \Omega_y\, \Omega_y^{-1},\, \Delta^{\,t} \rangle\!\rangle \\
& - \frac{1}{2} \langle\!\langle X\, X^{\,t},\, \Delta^{\,t}\, \Omega_y^{-1}\, \Delta \rangle\!\rangle - \frac{n}{2} \langle\!\langle L,\, \Delta^{\,t}\, \Omega_y^{-1}\, \Delta \rangle\!\rangle^{\beta} \\
= \; & \text{cste} + \frac{n}{2} \ln\det(\Omega_y) - \frac{n}{2} \langle\!\langle S_y, \Omega_y \rangle\!\rangle - n \langle\!\langle S_{yx}, \Delta \rangle\!\rangle - \frac{n}{2} \langle\!\langle S_x, \Delta^{\,t}\, \Omega_y^{-1}\, \Delta \rangle\!\rangle \\
& - \frac{n}{2} \langle\!\langle L,\, \Delta^{\,t}\, \Omega_y^{-1}\, \Delta \rangle\!\rangle^{\beta}
\end{aligned}
$$

The objective function is obtained by considering the negative log-likelihood on which we add the parameters and the penalty functions. The objective function being defined, we can now study the existence of a global minimum on its domain.

Proposition 2.2.2. *Assume that $\beta \geqslant 1$. Then, $L_{pa}(\Omega_y, \Delta)$ defined in (4) is jointly convex with respect to (Ω_y, Δ).*

Proof. See Section 2.4.2. □

Thus, for $\beta \geqslant 1$ the estimator of our model corresponds to the global minimum defined by

$$\widehat{\theta} = \arg\min_{\Theta} L_{\lambda,\mu,\eta}(\theta). \tag{5}$$

2.3 Theoretical guarantees

In this section provide some theoretical guarantees for the estimation of $\widehat{\theta}$ defined in (5). These are valid for any $\beta \geqslant 1$ and $\beta = 0$, and under the respect of the assumptions ($\mathrm{H}_{2.1}$) and ($\mathrm{H}_{2.2}$) which we will be discussed later in the chapter. However, we shall not theoretically deviate too much from the Gaussianity in the prior (*i.e.* $\beta = 1 + \epsilon$ for a small $\epsilon \geqslant 0$), even if we will allow ourselves some exceptions in the practical works. We will also come back to this point in due course.

Now and throughout the rest of the chapter, denote by $\theta = (\Omega_y, \Delta) \in \Theta = \mathbb{S}^q_{++} \times \mathbb{R}^{q \times p}$ the $(q \times (q+p))$-matrix of parameters of the model, with true value $\theta^* = (\Omega_y^*, \Delta^*)$. As it is usually done in studies implying sparsity, we will also consider S of cardinality $|S|$, the true active set of θ^* defined as $S = \{(i,j),\, \theta^*_{i,j} \neq 0\}$, and its complement $\bar{S}$. Theorem 2.3.1 gives an upper bound on the estimation error. To facilitate reading, we centralized the precise definition of the numerous constants involved and that of the assumptions to be verified in Section 2.4.3.

Theorem 2.3.1. *Fix $d_\lambda > c_\lambda > 1$, $d_\mu > c_\mu > 1$, $e_\lambda > 0$ and $e_\mu > 0$, and assume that the regularization parameters satisfy $(\lambda, \mu, \eta) \in \Lambda = [c_\lambda\, h_a,\, d_\lambda\, h_a] \times [c_\mu\, h_b,\, d_\mu\, h_b] \times [0,\, \overline{\eta}]$, where*

$$\overline{\eta} = \frac{\min\left\{ \frac{(c_\lambda - 1)\,\lambda}{c_\lambda\, \ell_a},\; \frac{(c_\mu - 1)\,\mu}{c_\mu\, \ell_b},\; \frac{e_\lambda\, h_a}{\ell_a},\; \frac{e_\mu\, h_b}{\ell_b} \right\}}{\beta\, s_L^{\beta - 1}}$$

for some non-random constants s_L, ℓ_a and ℓ_b defined in (10) *and* (11)*, and the random constants h_a and h_b given above. Then, under* ($\mathrm{H}_{2.1}$)*, there exists absolute constants $b_1 > 0$ and $b_2 > 0$ such that, for any $0 < b_3 < 1$ and as soon as $n > n_0$, with probability no less that $1 - \mathrm{e}^{-b_2 n} - b_3$, the estimator* (5) *satisfies*

$$\|\widehat{\theta} - \theta^*\|_F \leqslant \frac{16\, m^*\, c_{\lambda,\mu}\sqrt{|S|}}{\gamma_{r,\eta,\beta,p}} \sqrt{\frac{\ln(10(p+q)^2) - \ln(b_3)}{n}}$$

where $\gamma_{r,\eta,\beta,p}$, $c_{\lambda,\mu}$ and m^* are technical constants defined in (15)*,* (16) *and* (17)*, respectively, and where the minimal number of observations is given by*

$$\begin{aligned} n_0 \quad = \quad \max \Bigg\{ & \frac{(\ln(10(p+q)^2) - \ln(b_3))\, c_{\lambda,\mu}^2\, |S|\, (16\, m^*)^2}{r^{*\,2}\, \gamma_{r,\eta,\beta,p}^2}, \\ & b_1\, (q + \lceil s_\alpha \rceil \ln(p+q)), \ln(10(p+q)^2) - \ln(b_3) \Bigg\} \end{aligned} \tag{6}$$

with s_α defined in (13) *and r^* in* (14)*.*

Proof. See Section 2.4.3. □

Among all these constants, we can note that s_L, ℓ_a, ℓ_b, h_a and h_b are useful to properly describe and restrict Λ, the domain of validity of (λ, μ, η) for the theorem to hold. Once Λ is fixed, the other constants take part in the upper bound of the estimation error. However, as it stands, the theorem is very difficult to interpret. The next two remarks seem essential to have an overview of the orders of magnitude involved for the number of observations, for p and q, for the estimation error and for the regularization parameters.

Remark 2.3.1 (Validity band)**.** Of course the degree of sparsity $|S|$ is crucial in the estimation error, but it also plays an indirect role in the probability associated with the theorem and in the numerous constants. In virtue of Lemma 2.4.13, we can hope that λ and μ have a wide validity band, by playing on c_λ, c_μ, d_λ and d_μ. In turn, η also has a non-negligible area of validity, provided of course that ℓ_a, ℓ_b and s_L, all depending on combinations between Δ^*, $\Omega_y^{*\,-1}$ and L, are small enough. Accordingly, it would be to our advantage if L was both sparse and not chosen with too large elements. As it always appears together with η, we may as well take a normalized version of L (*e.g.* $|L|_\infty \leqslant 1$).

Remark 2.3.2 (Order of magnitude)**.** Even if the result holds for any $\beta \geqslant 1$, the terms $\propto p^{\beta-1}$ appearing in some upper bounds of the proof clearly argue in favor of a moderate choice $\beta \in [1, 1+\epsilon]$ for a small $\epsilon > 0$, depending on p. In other words, we cannot deviate too much from the Gaussianity in the prior on the direct links. For example in a very high-dimensional setting ($p \sim 10^7$), choosing $\epsilon = 0.1$ leads to $p^{\beta-1} \approx 5$ whereas we may try larger values of ϵ for the more common high-dimensional settings $p \sim 10^3$ or $p \sim 10^4$. By contrast, we can see that n_0 must (at least) grow like q for the theorem to hold, so high-dimensional responses are excluded. However in multiple-output regressions, even when p is extremely large, q generally remains small. According to all these considerations, we may roughly say that, in a high-dimensional setting with respect to p,

$$\|\widehat{\theta} - \theta^*\|_F \lesssim \sqrt{\frac{|S|\, \ln p}{n}}$$

with a large probability, under a suitable regularization of the model. We recognize the usual terms appearing in the error bounds of regressions with high-dimensional covariates, like the ℓ_2 error of the Lasso (see *e.g.* Chap. 11 of [HTW15]). This is the same bound as in [YZ14], but our additional structural penalty restricts Λ.

2.4 Technical proofs

This proof section is organized in three parts. First, we introduce some useful linear algebra lemmas that will be repeatedly used subsequently, well-known for most of them. We then prove the joint convexity of the objective function (4), and finish with the proof of our main result.

2.4.1 Linear algebra

Lemma 2.4.1. *Let $A \in \mathbb{S}_+^d$ and $U \in \mathbb{R}^{d\times\ell}$. Then, $U^t A U \in \mathbb{S}_+^\ell$.*

Proof. Since A is symmetric with non-negative eigenvalues, there is an orthogonal matrix P such that $A = PDP^t$ with $D = \mathrm{diag}(\mathrm{sp}(A)) \in \mathbb{S}_+^d$. Thus, for all $v \in \mathbb{R}^\ell$, it follows that

$$\langle v, U^t A\, U\, v\rangle \;=\; v^t\, U^t P D P^t\, U\, v \;=\; \|D^{1/2}\, P^t\, U\, v\|^2 \geqslant 0.$$

□

Lemma 2.4.2. *Let $A \in \mathbb{S}_{++}^d$ and $B \in \mathbb{S}_+^d$. Then for all $i \in \{1, ..., d\}$, $\lambda_i(AB) \geqslant 0$.*

Proof. The equality $AB = A^{1/2}\,(A^{1/2}\,B\,A^{1/2})\,A^{-1/2}$ shows that AB and $A^{1/2}\,B\,A^{1/2}$ are similar, so they must share the same eigenvalues. From Lemma 2.4.1, $\lambda_i(A^{1/2}\,B\,A^{1/2}) \geqslant 0$. □

Lemma 2.4.3. *Let $A \in \mathbb{S}_+^d$ and $B \in \mathbb{S}_+^d$. Then,*

$$\lambda_{\min}(A)\,\mathrm{tr}(B) \;\leqslant\; \mathrm{tr}(AB) \;\leqslant\; \lambda_{\max}(A)\,\mathrm{tr}(B).$$

Proof. Since $A - \lambda_{\min}(A) I_d \in \mathbb{S}_+^d$ and $B \in \mathbb{S}_+^d$,

$$\mathrm{tr}((A - \lambda_{\min}(A) I_d)\, B) = \mathrm{tr}(B^{1/2}\,(A - \lambda_{\min}(A) I_d)\, B^{1/2}) \geqslant 0$$

from Lemma 2.4.1, thus $\mathrm{tr}(AB) \geqslant \lambda_{\min}(A)\,\mathrm{tr}(B)$. The other inequality is obtained through $\lambda_{\max}(A) I_d - A \in \mathbb{S}_+^d$. □

Lemma 2.4.4. *Let $A \in \mathbb{S}_{++}^d$ and $B \in \mathbb{S}_+^d$. Then,*

$$\lambda_{\min}(A)\,\lambda_{\min}(B) \;\leqslant\; \lambda_{\min}(AB) \quad \textit{and} \quad \lambda_{\max}(AB) \;\leqslant\; \lambda_{\max}(A)\,\lambda_{\max}(B).$$

Proof. On the one hand,

$$\lambda_{\max}(AB) \;\leqslant\; \|AB\|_2 \;\leqslant\; \|A\|_2\,\|B\|_2 \;=\; \lambda_{\max}(A)\,\lambda_{\max}(B),$$

since A and B are symmetric and since, from Lemma 2.4.2 and by hypothesis, all eigenvalues appearing in the relation are non-negative. Suppose now that B is invertible so that both A^{-1} and B^{-1} belong to $\mathbb{S}_{++}^d$. Then,

$$\lambda_{\max}((AB)^{-1}) \leqslant \lambda_{\max}(A^{-1})\,\lambda_{\max}(B^{-1}) \qquad \Longleftrightarrow \qquad \lambda_{\min}(AB) \geqslant \lambda_{\min}(A)\,\lambda_{\min}(B).$$

However, if B is not invertible, the relation trivially holds since we still have $\lambda_{\min}(AB) \geqslant 0$ from Lemma 2.4.2. □

Lemma 2.4.5. *Let $A \in \mathbb{S}_+^d$ and $U \in \mathbb{R}^{d\times\ell}$. Then,*

$$\lambda_{\min}(A)\,\|U\|_F^2 \leqslant \operatorname{tr}(U^{\,t}AU) \leqslant \lambda_{\max}(A)\,\|U\|_F^2.$$

Proof. Denote by u_i the i-th column of U, and let P be the orthogonal matrix such that $A = PDP^t$ with $D = \operatorname{diag}(\operatorname{sp}(A)) \in \mathbb{S}_+^d$. The i-th diagonal element of $U^{\,t}AU$ satisfies $u_i^t A\,u_i = u_i^t PDP^t\,u_i \geqslant \lambda_{\min}(A)\,\|u_i\|^2 \geqslant 0$. Thus,

$$\operatorname{tr}(U^{\,t}AU) = \sum_{i=1}^{\ell} u_i^t A\,u_i \geqslant \lambda_{\min}(A)\sum_{i=1}^{\ell}\|u_i\|^2 = \lambda_{\min}(A)\,\|U\|_F^2.$$

The upper bound stems from $0 \leqslant u_i^t A\,u_i \leqslant \lambda_{\max}(A)\,\|u_i\|^2$. □

Lemma 2.4.6. *Let A and B be symmetric matrices of same dimensions. Then,*

$$\lambda_{\min}(A) + \lambda_{\min}(B) \leqslant \lambda_{\min}(A+B) \quad \text{and} \quad \lambda_{\max}(A+B) \leqslant \lambda_{\max}(A) + \lambda_{\max}(B).$$

Proof. These are just two special cases of Weyl inequalities. We refer the reader to Theorem. 4.3.1 of [HJ12], for example. □

2.4.2 Convexity of the objective function

We know from Prop. 1 of [YZ14] and the convexity of the elementwise ℓ_1 norm that $L_{pa}(\Omega_y, \Delta) - \eta\langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\Delta\rangle\!\rangle^{\beta}$ is itself convex, but it remains to show that this is still the case with the additional smooth penalty.

Proof of Proposition 2.2.2

We want to show that $\langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\Delta\rangle\!\rangle^{\beta}$ satisfies the convexity inequality. Recall that $\Theta = \mathbb{S}_{++}^q \times \mathbb{R}^{q\times p}$ and consider the mapping $\Phi : \Theta \to \mathbb{S}_+^p$ defined as

$$\forall\,(A,B) \in \Theta, \qquad \Phi(A,B) = B^{\,t}A^{-1}B.$$

We can already note from Lemma 2.4.1 that $\operatorname{tr}(\Phi(A,B)) \geqslant 0$. Moreover, for all $0 \leqslant h \leqslant 1$ and all $Z_i = (A_i, B_i) \in \Theta$, $i = 1, 2$, we are interested in the matrix $M_h(Z_1, Z_2)$ satisfying the following blockwise

decomposition

$$M_h(Z_1, Z_2) = h \begin{pmatrix} A_1 & B_1 \\ B_1^t & B_1^t\, A_1^{-1} B_1 \end{pmatrix} + (1-h) \begin{pmatrix} A_2 & B_2 \\ B_2^t & B_2^t\, A_2^{-1} B_2 \end{pmatrix}. \tag{7}$$

The Schur complement $S_h(Z_1, Z_2)$ of block $hA_1 + (1-h)A_2$ in $M_h(Z_1, Z_2)$ is defined by

$$\begin{aligned} S_h(Z_1, Z_2) \;&=\; h\,(B_1^t\, A_1^{-1}\, B_1) + (1-h)\,(B_2^t\, A_2^{-1}\, B_2) \\ &\quad - (h\, B_1^t + (1-h)\, B_2^t)\,(h\, A_1 + (1-h)\, A_2)^{-1}\,(h\, B_1 + (1-h)\, B_2), \\ &=\; h\,\Phi(Z_1) + (1-h)\,\Phi(Z_2) - \Phi(hZ_1 + (1-h)Z_2). \end{aligned} \tag{8}$$

Moreover, the decomposition

$$\begin{pmatrix} A^{1/2} & A^{-1/2}\, B \\ 0 & 0 \end{pmatrix}^t \begin{pmatrix} A^{1/2} & A^{-1/2}\, B \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} A & B \\ B^t & B^t\, A^{-1} B \end{pmatrix}$$

directly shows that $M_h(Z_1, Z_2)$ in (7) is symmetric and positive semi-definite. It is well-known (see *e.g.* Appendix A.5.5 of [BBV04]) that in that case, the Schur complement (8) must also be positive semi-definite. The trace of $S_h(Z_1, Z_2)$ is therefore positive. For $i = 1, 2$ we pose $Z_i = (\Omega_{i,yy}, \Omega_{i,yx} L^{1/2})$, $P_h = h\,\Delta_1 + (1-h)\,\Delta_2$, $Q_h = h\,\Omega_{1,yy} + (1-h)\,\Omega_{2,yy}$ and $\beta \geqslant 1$, we obtain

$$\begin{aligned} \langle\!\langle L, P_h^t\, Q_h^{-1} P_h \rangle\!\rangle^\beta \;&=\; (\mathrm{tr}(\Phi(hZ_1 + (1-h)Z_2)))^\beta \\ &\leqslant\; (h\,\mathrm{tr}(\Phi(Z_1)) + (1-h)\,\mathrm{tr}(\Phi(Z_2)))^\beta \\ &=\; (h\,\langle\!\langle L, \Delta_1^t\, \Omega_{1,yy}^{-1}\, \Delta_1 \rangle\!\rangle + (1-h)\,\langle\!\langle L, \Delta_2^t\, \Omega_{2,yy}^{-1}\, \Delta_2 \rangle\!\rangle)^\beta \\ &\leqslant\; h\,\langle\!\langle L, \Delta_1^t\, \Omega_{1,yy}^{-1}\, \Delta_1 \rangle\!\rangle^\beta + (1-h)\,\langle\!\langle L, \Delta_2^t\, \Omega_{2,yy}^{-1}\, \Delta_2 \rangle\!\rangle^\beta \end{aligned}$$

This convexity inequality concludes the proof.

2.4.3 Theoretical guarantees

Before stating the proof of the theorem, we first present the assumptions related to the covariances ; and then we define the constants appearing in its statement, and which will be used throughout its proof.

Assumptions

Our results depend on some basic assumptions related to the true covariances of the Gaussian observations and the empirical covariances.

$$\Sigma_x^* \in \mathbb{S}_{++}^p, \qquad \Omega_y^* \in \mathbb{S}_{++}^q, \qquad B \neq 0 \text{ (that is, } \Delta^* \neq 0) \qquad \text{and} \qquad \Delta^*\, L\, \Delta^{*\,t} \in \mathbb{S}_{++}^q. \tag{$\mathrm{H}_{2.1}$}$$

($\mathrm{H}_{2.1}$) is a natural hypothesis in our framework, in particular we suppose that there is at least a link between X and Y.

Remark 2.4.7 (Null model)**.** Even if it is of less interest, our study does not exclude the case where $\Delta^* = 0$. Indeed, we might as well consider that $\Delta^* = 0$ and get the same results, but some constants

should be refined. On the other hand, $\Sigma_x^* \in \mathbb{S}_{++}^p$ and $\Omega_y^* \in \mathbb{S}_{++}^q$ are crucial.

$$\forall\, u \neq 0 \text{ tel que } |u|_0 \leqslant \lceil s_\alpha \rceil, \qquad \frac{1}{2}\, u^{\,t}\, \Sigma_x^*\, u \;\leqslant\; u^{\,t}\, S_x\, u \;\leqslant\; \frac{3}{2}\, u^{\,t}\, \Sigma_x^*\, u. \tag{$\mathrm{H}_{2.2}$}$$

$$\text{De plus,} \qquad \lambda_{\max}(\Delta^*\, S_x\, \Delta^{*\,t}) \;\leqslant\; \frac{7}{5}\, \lambda_{\max}(\Delta^*\, \Sigma_x^*\, \Delta^{*\,t}).$$

($\mathrm{H}_{2.2}$) is to be assumed with the smallest integer greater than s_α in (13). This is a random hypothesis, which will be controlled with a probability, related to the proximity between the empirical covariance and the true covariance of the predictors, since we recall that S has no reason to be an excellent approximation of Σ^* when $p \gg n$. This is also assumed by the authors of [YZ14], it is a kind of restricted isometry propertie (RIP), well-known in high-dimensional studies. In particular, we will see through Lemma 2.4.13 that it is satisfied with high probability provided that n is large enough.

Some constants

First we define the random constants. Under ($\mathrm{H}_{2.1}$), the random matrices

$$A_n = (S_y - \Sigma_y^*) - {\Omega_y^*}^{-1}\, \Delta^*\, (S_x - \Sigma_x^*)\, \Delta^{*\,t}\, {\Omega_y^*}^{-1} \qquad \text{with} \qquad h_a = |A_n|_\infty \tag{$\mathrm{C}_{2.1}$}$$

and

$$B_n = 2\, ((S_{yx} - \Sigma_{yx}^*) + {\Omega_y^*}^{-1}\, \Delta^*\, (S_x - \Sigma_x^*)) \qquad \text{with} \qquad h_b = |B_n|_\infty \tag{$\mathrm{C}_{2.2}$}$$

are going to play a fundamental role, especially h_a and h_b.

Let now define the non-random constants, starting with those related to L and the true values of the model. The bounds

$$\underline{\omega}_L = \frac{\lambda_{\min}(\Delta^*\, L\, \Delta^{*\,t})}{4\, \lambda_{\max}(\Omega_y^*)}, \qquad \overline{\omega}_L = \frac{4\, \lambda_{\max}(\Delta^*\, L\, \Delta^{*\,t})}{\lambda_{\min}(\Omega_y^*)}, \qquad \overline{\omega}_S = \frac{4\, \lambda_{\max}(\Delta^*\, \Sigma_x^*\, \Delta^{*\,t})}{\lambda_{\min}(\Omega_y^*)}. \tag{9}$$

are useful to control the eigenvalues of some recurrent expressions (Lemmas 2.4.8 and 2.4.9), uniformly in a neighborhood of $\theta^* = (\Omega_y^*, \Delta^*)$. The true value of the term at the heart of the structural regularization is

$$s_L = \langle\!\langle L, \Delta^{*\,t}\, {\Omega_y^*}^{-1}\, \Delta^* \rangle\!\rangle. \tag{10}$$

It plays a role in the proof of Lemma 2.4.10 and, as a consequence, in the definition of the area of validity Λ. This important lemma also requires to define

$$\ell_a = |{\Omega_y^*}^{-1}\, \Delta^*\, L\, \Delta^{*\,t}\, {\Omega_y^*}^{-1}|_\infty \qquad \text{and} \qquad \ell_b = 2\, |{\Omega_y^*}^{-1}\, \Delta^*\, L|_\infty \tag{11}$$

and, in the context of the theorem,

$$\alpha = \frac{\max\left\{ \frac{(c_\lambda+1)\lambda}{c_\lambda} + \eta\beta\, s_L^{\beta-1}\ell_a,\, \frac{(c_\mu+1)\mu}{c_\mu} + \eta\beta\, s_L^{\beta-1}\ell_b \right\}}{\min\left\{ \frac{(c_\lambda-1)\lambda}{c_\lambda} - \eta\beta\, s_L^{\beta-1}\ell_a,\, \frac{(c_\mu-1)\mu}{c_\mu} - \eta\beta\, s_L^{\beta-1}\ell_b \right\}}. \tag{12}$$

From α and the cardinality of the true active set $|S|$, let

$$s_\alpha = |S| \left[1 + \frac{12\,\alpha^2\,\lambda_{\max}(\Sigma_x^*)}{\lambda_{\min}(\Sigma_x^*)} \right] \tag{13}$$

which serves as an upper bound in the random hypothesis ($\mathrm{H}_{2.2}$). Similarly, let

$$r^* = \min\{r_1^*, r_2^*, r_3^*, r_4^*\} \tag{14}$$

where

$$r_1^* = \frac{\lambda_{\min}(\Omega_y^*)}{2}, \qquad r_2^* = \frac{\frac{\sqrt{10}-\sqrt{7}}{\sqrt{5}}\sqrt{\lambda_{\max}(\Delta^*\,\Sigma_x^*\,\Delta^{*\,t})}}{\frac{3\sqrt{3}}{2\sqrt{2}}\sqrt{\lambda_{\max}(\Sigma_x^*)}}, \qquad r_3^* = \frac{\lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t})}{4\,\|L\,\Delta^{*\,t}\|_2}$$

and

$$r_4^* = \frac{(\sqrt{2}-1)\sqrt{\lambda_{\max}(\Delta^*\,L\,\Delta^{*\,t})}}{\sqrt{\lambda_{\max}(L)}}.$$

Together with α given above, r^* is necessary to build the so-called neighborhood $N_{r,\alpha}(\theta^*)$ defined in (22), which plays a fundamental role in all our reasonings. It is important to note that, under the configuration of the theorem and hypothesis ($\mathrm{H}_{2.1}$), $\alpha > 0$ and $r^* > 0$. Then, Lemma 2.4.11 highlights a new constant, characterizing a strong local convexity of the smooth part of the objective in the neighborhood $N_{r,\alpha}(\theta^*)$,

$$\gamma_{r,\eta,\beta,p} = \min\left\{ \frac{a_1}{8\,\lambda_{\max}^2(\Omega_y^*)},\ \frac{a_2\,\lambda_{\min}(L)}{4\,\lambda_{\max}(\Omega_y^*)} + \frac{a_3\,\lambda_{\min}(\Sigma_x^*)}{40\,\lambda_{\max}(\Omega_y^*)} \right\} \tag{15}$$

where, as it is detailed in the proof of the lemma in question,

$$a_1 = 1 - \epsilon_S\,\overline{\omega}_S - \eta\beta\,p^{\,\beta-1}\,\overline{\omega}_L^{\,\beta}\,\epsilon_L, \qquad a_2 = \frac{2\,\epsilon_S}{2+\epsilon_S} \qquad \text{and} \qquad a_3 = \eta\beta\,(p\,\underline{\omega}_L)^{\beta-1}\,\frac{2\,\epsilon_L}{2+\epsilon_L}$$

for some well-chosen $\epsilon_S > 0$ and $\epsilon_L > 0$. Here again, we make sure that $\gamma_{r,\eta,\beta,p} > 0$. In the same way, in the context of the theorem,

$$c_{\lambda,\mu} = \max\left\{ \frac{(c_\lambda+1)\,d_\lambda}{c_\lambda} + e_\lambda,\ \frac{(c_\mu+1)\,d_\mu}{c_\mu} + e_\mu \right\} \tag{16}$$

is needed through Lemma 2.4.12. Finally, independently of the structure matrix L,

$$m^* = |\mathrm{diag}(\Sigma_x^*)|_\infty + |\mathrm{diag}(\Omega_y^{*\,-1}\,\Delta^*\,\Sigma_x^*\,\Delta^{*\,t}\,\Omega_y^{*\,-1})|_\infty \tag{17}$$

is going to play a significative role in the upper bound of the theorem.

Proof of Theorem 2.3.1

Let $R_n(\theta)$ be the the smooth part of the objective (4),

$$\begin{aligned} R_n(\theta) \quad = \quad & -\ln\det(\Omega_y) + \langle\!\langle S_y, \Omega_y \rangle\!\rangle + 2\,\langle\!\langle S_{yx}, \Delta \rangle\!\rangle \\ & + \langle\!\langle S_x, \Delta^{\,t}\,\Omega_y^{-1}\,\Delta \rangle\!\rangle + \eta\,\langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\Delta \rangle\!\rangle^\beta. \end{aligned} \tag{18}$$

For any $\theta \in \Theta$ and $t \in \mathbb{R}$, by a Taylor expansion,

$$R_n(\theta^* + t\,(\theta - \theta^*)) = R_n(\theta^*) + t\,\langle\!\langle \nabla R_n(\theta^*), \theta - \theta^* \rangle\!\rangle + e_t(\theta, \theta^*) \tag{19}$$

for some second-order error term $e_t(\theta, \theta^*)$. Consider the reparametrization

$$\phi(t) = R_n(\theta^* + t\,(\theta - \theta^*)) \tag{20}$$

so that $\phi'(0) = \langle\!\langle \nabla R_n(\theta^*), \theta - \theta^* \rangle\!\rangle$. Let $\delta\theta_y = \Omega_y - \Omega_y^*$ and $\delta\theta_{yx} = \Delta - \Delta^*$, let also $\delta\theta = \theta - \theta^*$ in a compact form. The estimation error is denoted

$$\delta\vartheta = \widehat{\theta} - \theta^* = (\widehat{\Omega}_y - \Omega_y^*, \widehat{\Delta} - \Delta^*) = (\delta\vartheta_y, \delta\vartheta_{yx}). \tag{21}$$

Before we start the actual proof, some additional lemmas are needed. They constitute a local study in a sort of r^*-neighborhood of θ^* that we define as

$$N_{r,\alpha}(\theta^*) = \big\{\theta \in \Theta,\ \|\delta\theta\|_F \leqslant r^* \text{ and } |[\delta\theta]_{\bar{S}}|_1 \leqslant \alpha |[\delta\theta]_S|_1\big\}. \tag{22}$$

Our strategy can be summarized as follows:

→ (Lemma 2.4.10) Show that there exists a configuration for the regularization parameters (λ, μ, η) so that the estimation error satisfies $|[\delta\vartheta]_{\bar{S}}|_1 \leqslant \alpha |[\delta\vartheta]_S|_1$ for some $\alpha > 0$.

→ (Lemma 2.4.11) Find some $r^* > 0$ and $\gamma_{r,\eta,\beta,p} > 0$ such that $e_1(\theta, \theta^*) > \gamma_{r,\eta,\beta,p} \|\delta\theta\|_F^2$ as soon as $\theta \in N_{r,\alpha}(\theta^*)$.

→ (Lemma 2.4.12) Exploit this result to show that the estimation error must also satisfy $\|\delta\vartheta\|_F \leqslant r^*$ provided that $\max\{h_a, h_b\}$ is small enough.

→ (Lemma 2.4.13) Conclude that the theorem holds with high probability, provided that n is large enough.

Thereafter, $N_{r,\alpha}(\theta^*)$ will always refer to α in (12) and r^* in (14). The next two lemmas give some bounds for expressions that will appear repeatedly.

Lemma 2.4.8. *Under* ($\mathrm{H}_{2.1}$) *and* ($\mathrm{H}_{2.2}$), *for all* $\theta \in N_{r,\alpha}(\theta^*)$, *we have the bound*

$$\lambda_{\max}(\Omega_y^{-1}\,\Delta\,S_x\,\Delta^{\,t}) \leqslant \overline{\omega}_S$$

where $\overline{\omega}_S$ *is given in* (9). *In addition,*

$$\mathrm{tr}(\delta\theta_{yx}\,S_x\,\delta\theta_{yx}^t) \geqslant \frac{\lambda_{\min}(\Sigma_x^*)}{10}\,\|\delta\theta_{yx}\|_F^2.$$

Proof. Similar reasonings may be found in the proofs of Lemmas 1-2 of [YZ14]. We simply reworked the constants to make them stick to our study. □

Lemma 2.4.9. *Under* ($\mathrm{H}_{2.1}$)*, for all* $\theta \in N_{r,\alpha}(\theta^*)$*, we have the bounds*

$$\lambda_{\min}(\Omega_y^{-1}\,\Delta\,L\,\Delta^{\,t}) \geqslant \underline{\omega}_L \quad \textit{and} \quad \lambda_{\max}(\Omega_y^{-1}\,\Delta\,L\,\Delta^{\,t}) \leqslant \overline{\omega}_L$$

where $\underline{\omega}_L$ *and* $\overline{\omega}_L$ *are given in* (9)*. As a corollary,*

$$p\,\underline{\omega}_L \leqslant \langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\Delta \rangle\!\rangle \leqslant p\,\overline{\omega}_L.$$

Proof. Let $\theta = \theta^* + \delta\theta \in N_{r,\alpha}(\theta^*)$, we have $\Delta\,L\,\Delta^{\,t} = (\Delta^* + \delta\theta_{yx})\,L\,(\Delta^{*\,t} + \delta\theta_{yx}^t)$. Since $L \in \mathbb{S}_{++}^p$ from Lemma 2.4.1, $\delta\theta_{yx}\,L\,\delta\theta_{yx}^t \in \mathbb{S}_+^q$, and in particular $\lambda_{\min}(\delta\theta_{yx}\,L\,\delta\theta_{yx}^t) \geqslant 0$. From Lemma 2.4.6,

$$\begin{aligned}
2\,\lambda_{\min}(\Delta\,L\,\Delta^{\,t}) &\geqslant 2\left(\lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t}) + \lambda_{\min}(\delta\theta_{yx}\,L\,\Delta^{*\,t} + \Delta^*\,L\,\delta\theta_{yx}^t)\right) + 2\,\lambda_{\min}(\delta\theta_{yx}\,L\,\delta\theta_{yx}^t) \\
&\geqslant 2\left(\lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t}) + \lambda_{\min}(\delta\theta_{yx}\,L\,\Delta^{*\,t} + \Delta^*\,L\,\delta\theta_{yx}^t)\right) \\
&\geqslant 2\left(\lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t}) - \|\delta\theta_{yx}\,L\,\Delta^{*\,t} + \Delta^*\,L\,\delta\theta_{yx}^t\|_2\right) \\
&\geqslant 2\left(\lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t}) - \|\delta\theta_{yx}\,L\,\Delta^{*\,t}\|_2 - \|\Delta^*\,L\,\delta\theta_{yx}^t\|_2\right) \\
&\geqslant 2\left(\lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t}) - 2\,\|\delta\theta_{yx}\|_2\,\|L\,\Delta^{*\,t}\|_2\right) \\
&\geqslant 2\left(\lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t}) - 2\,\|\delta\theta_{yx}\|_F\,\|L\,\Delta^{*\,t}\|_2\right) \\
&\geqslant \lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t})
\end{aligned}$$

as soon as $\|\delta\theta_{yx}\|_F \leqslant r^*$ since we know that $4\,\|L\,\Delta^{*\,t}\|_2\,r^* \leqslant \lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t})$. We therefore deduce that $\Delta\,L\,\Delta^{\,t} \in \mathbb{S}_{++}^q$. From Lemma 2.4.4, we get

$$\lambda_{\min}(\Omega_y^{-1}\,\Delta\,L\,\Delta^{\,t}) \geqslant \frac{\lambda_{\min}(\Delta\,L\,\Delta^{\,t})}{\lambda_{\max}(\Omega_y)} \geqslant \frac{\lambda_{\min}(\Delta^*\,L\,\Delta^{*\,t})}{4\,\lambda_{\max}(\Omega_y^*)}$$

where the inequality in the denominator comes from $\lambda_{\max}(\Omega_y) \leqslant \lambda_{\max}(\Omega_y^*) + \lambda_{\max}(\delta\theta_y)$, *via* Lemma 2.4.6, and the fact that $\lambda_{\max}(\delta\theta_y) \leqslant \|\delta\theta_y\|_F \leqslant r^* \leqslant \lambda_{\max}(\Omega_y^*)$. For the upper bound, a similar logic gives, with Lemmas 2.4.6 and 2.4.5,

$$\begin{aligned}
\sqrt{\lambda_{\max}(\Delta\,L\,\Delta^{\,t})} &= \|(\Delta^* + \delta\theta_{yx})\,L^{\,1/2}\|_2 \\
&\leqslant \|\Delta^*\,L^{\,1/2}\|_2 + \|\delta\theta_{yx}\,L^{\,1/2}\|_2 \\
&\leqslant \sqrt{\lambda_{\max}(\Delta^*\,L\,\Delta^{*\,t})} + \sqrt{\mathrm{tr}(\delta\theta_{yx}\,L\,\delta\theta_{yx}^t)} \\
&\leqslant \sqrt{\lambda_{\max}(\Delta^*\,L\,\Delta^{*\,t})} + \|\delta\theta_{yx}\|_F\,\sqrt{\lambda_{\max}(L)} \\
&\leqslant \sqrt{2\,\lambda_{\max}(\Delta^*\,L\,\Delta^{*\,t})}
\end{aligned}$$

since $\|\delta\theta_{yx}\|_F \leqslant r^*$ and $r^*\sqrt{\lambda_{\max}(L)} \leqslant (\sqrt{2}-1)\sqrt{\lambda_{\max}(\Delta^*\,L\,\Delta^{*\,t})}$. It follows from Lemma 2.4.4 that

$$\lambda_{\max}(\Omega_y^{-1}\,\Delta\,L\,\Delta^{\,t}) \leqslant \frac{\lambda_{\max}(\Delta\,L\,\Delta^{\,t})}{\lambda_{\min}(\Omega_y)} \leqslant \frac{4\,\lambda_{\max}(\Delta^*\,L\,\Delta^{*\,t})}{\lambda_{\min}(\Omega_y^*)}$$

where the inequality in the denominator comes from $\lambda_{\min}(\Omega_y) \geqslant \lambda_{\min}(\Omega_y^*) + \lambda_{\min}(\delta\theta_y)$, *via* Lemma 2.4.6, and the fact that $2\,\lambda_{\min}(\delta\theta_y) \geqslant -2\,\|\delta\theta_y\|_F \geqslant -2\,r^* \geqslant -\lambda_{\min}(\Omega_y^*)$. The corollary that concludes the lemma

is now immediate. □

Lemma 2.4.10. *Assume that λ, μ and η are chosen according to the configuration of the theorem. Then, under* ($\mathrm{H}_{2.1}$)*, the estimation error satisfies*

$$|[\delta\vartheta]_{\bar{S}}|_1 \leqslant \alpha\,|[\delta\vartheta]_S|_1$$

where $\alpha > 0$ is given in (12)*.*

Proof. Taking $t = 1$ in the Taylor expansion (19) with $\theta = \widehat{\theta}$ and considering the definition of ϕ in (20), by convexity,

$$R_n(\widehat{\theta}) - R_n(\theta^*) \geqslant \phi'(0) = \langle\!\langle \nabla R_n(\theta^*), \theta - \theta^* \rangle\!\rangle.$$

The first derivative of ϕ will be explicitely computed in (26). For $t = 0$, we find

$$\begin{aligned}
\nabla R_n(\theta^*) &= -(\Omega_y^{*\,-1},\, 0_{q,p}) + (S_y,\, 0_{q,p}) + 2\,(0_{q,q},\, S_{yx}) + (-\Omega_y^{*\,-1}\,\Delta^*\,S_x\,\Delta^{*\,t}\,\Omega_y^{*\,-1},\, 2\,\Omega_y^{*\,-1}\,\Delta^*\,S_x) \\
&\quad + \eta\,\beta\,\langle\!\langle L, \Delta^{*\,t}\,\Omega_y^{*\,-1}\,\Delta^* \rangle\!\rangle^{\beta-1}\,(-\Omega_y^{*\,-1}\,\Delta^*\,L\,\Delta^{*\,t}\,\Omega_y^{*\,-1},\, 2\,\Omega_y^{*\,-1}\,\Delta^*\,L), \\
\phi'(0) &= -\langle\!\langle \Omega_y^{*\,-1}, \delta\vartheta_y \rangle\!\rangle + \langle\!\langle S_y, \delta\vartheta_y \rangle\!\rangle + 2\,\langle\!\langle S_{yx}, \delta\vartheta_{yx} \rangle\!\rangle \\
&\quad - \langle\!\langle \Omega_y^{*\,-1}\,\Delta^*\,S_x\,\Delta^{*\,t}\,\Omega_y^{*\,-1}, \delta\vartheta_y \rangle\!\rangle + 2\,\langle\!\langle \Omega_y^{*\,-1}\,\Delta^*\,S_x, \delta\vartheta_{yx} \rangle\!\rangle \\
&\quad + \eta\beta\,\langle\!\langle L, \Delta^{*\,t}\,\Omega_y^{*\,-1}\,\Delta^* \rangle\!\rangle^{\beta-1}\,\big[-\langle\!\langle \Omega_y^{*\,-1}\,\Delta^*\,L\,\Delta^{*\,t}\,\Omega_y^{*\,-1}, \delta\vartheta_y \rangle\!\rangle + 2\,\langle\!\langle \Omega_y^{*\,-1}\,\Delta^*\,L, \delta\vartheta_{yx} \rangle\!\rangle \big].
\end{aligned}$$

Moreover, note that by using the blockwise relations (29), we can show that the random matrices A_n (with norm max h_a) and B_n (with norm max h_b) defined in ($\mathrm{C}_{2.1}$) and ($\mathrm{C}_{2.2}$) verify

$$\begin{aligned}
A_n &= S_y - \Omega_y^{*\,-1}\,\Delta^*\,S_x\,\Delta^{*\,t}\,\Omega_y^{*\,-1} - \Sigma_y^* + \Omega_y^{*\,-1}\,\Omega_y^*\,\Sigma_{yx}^*\,\Sigma_x^{*\,-1}\Sigma_x^*\,\Sigma_x^{*\,-1}\,\Sigma_{yx}^{*\,t}\,\Omega_y^*\,\Omega_y^{*\,-1} \\
&= -\Omega_y^{*\,-1} + S_y - \Omega_y^{*\,-1}\,\Delta^*\,S_x\,\Delta^{*\,t}\,\Omega_y^{*\,-1}
\end{aligned}$$

and

$$\begin{aligned}
B_n &= 2\,S_{yx} + 2\,\Omega_y^{*\,-1}\,\Delta^*\,S_x - 2\,\Sigma_{yx}^* + 2\,\Omega_y^{*\,-1}\,\Omega_y^*\,\Sigma_{yx}^*\,\Sigma_x^{*\,-1}\,\Sigma_x^* \\
&= 2\,S_{yx} + 2\,\Omega_y^{*\,-1}\,\Delta^*\,S_x.
\end{aligned}$$

By posing $C_A = -\Omega_y^{*\,-1}\,\Delta^*\,L\,\Delta^{*\,t}\,\Omega_y^{*\,-1}$ and $C_B = 2\,\Omega_y^{*\,-1}\,\Delta^*\,L$, and considering s_L given in (11), we obtain a compact form of $\phi'(0)$

$$\phi'(0) = \langle\!\langle A_n + \eta\beta\,s_L^{\beta-1} C_A, \delta\vartheta_y \rangle\!\rangle + \langle\!\langle B_n + \eta\beta\,s_L^{\beta-1} C_B, \delta\vartheta_{yx} \rangle\!\rangle.$$

Whence it follows from the well-known relation $|\mathrm{tr}(M_1 M_2)| \leqslant |M_1|_\infty\,|M_2|_1$, where M_1 and M_2 are compatible matrices, that

$$\begin{aligned}
\phi'(0) &\geqslant -h_a\,|\delta\vartheta_y|_1 - \eta\beta\,s_L^{\beta-1}\ell_a\,|\delta\vartheta_y|_1 - h_b\,|\delta\vartheta_{yx}|_1 - \eta\beta\,s_L^{\beta-1}\ell_b\,|\delta\vartheta_{yx}|_1 \\
&\geqslant -\frac{\lambda}{c_\lambda}\,|\delta\vartheta_y|_1 - \eta\beta\,s_L^{\beta-1}\ell_a\,|\delta\vartheta_y|_1 - \frac{\mu}{c_\mu}\,|\delta\vartheta_{yx}|_1 - \eta\beta\,s_L^{\beta-1}\ell_b\,|\delta\vartheta_{yx}|_1,
\end{aligned}$$

making use of the constants (11), $\lambda \geqslant c_\lambda\, h_a$ and $\mu \geqslant c_\mu\, h_b$. For the sake of clarity, let

$$\Delta_n(\theta, \theta^*) = R_n(\theta) + \lambda\,|\Omega_y|_1^- + \mu\,|\Delta|_1 - R_n(\theta^*) - \lambda\,|\Omega_y^*|_1^- - \mu\,|\Delta^*|_1.$$

For all $\theta \in \Theta$,

$$\begin{aligned} |\Omega_y|_1^- - |\Omega_y^*|_1^- &= |[\Omega_y^* + \delta\theta_y]_S|_1^- + |[\delta\theta_y]_{\bar{S}}|_1^- - |[\Omega_y^*]_S|_1^- \\ &\geqslant \big||[\Omega_y^*]_S|_1^- - |[\delta\theta_y]_S|_1^-\big| + |[\delta\theta_y]_{\bar{S}}|_1^- - |[\Omega_y^*]_S|_1^- \\ &\geqslant |[\delta\theta_y]_{\bar{S}}|_1 - |[\delta\theta_y]_S|_1 \end{aligned}$$

from the triangle inequality and the fact that, as Ω_y^* is positive definite, the diagonal must belong to S, *i.e.* $(j,j) \in S$ for all $1 \leqslant j \leqslant q$ so that any square matrix M of size q is such that $[M]_{\bar{S}}$ has diagonal elements all equal to zero. A similar bound obviously holds for $|\Delta|_1 - |\Delta^*|_1$. Now, using the previous information, we can show that

$$\begin{aligned} \Delta_n(\widehat{\theta}, \theta^*) &\geqslant -\Big(\frac{\lambda}{c_\lambda} + \eta\beta\, s_L^{\beta-1}\ell_a\Big)|\delta\vartheta_y|_1 - \Big(\frac{\mu}{c_\mu} + \eta\beta\, s_L^{\beta-1}\ell_b\Big)|\delta\vartheta_{yx}|_1 \\ &\quad + \lambda\big(|[\delta\theta_y]_{\bar{S}}|_1 - |[\delta\theta_y]_S|_1\big) + \mu\big(|[\delta\theta_{yx}]_{\bar{S}}|_1 - |[\delta\theta_{yx}]_S|_1\big), \\ &= \Big(-\frac{\lambda}{c_\lambda} - \eta\beta\, s_L^{\beta-1}\ell_a + \lambda\Big)|[\delta\theta_y]_{\bar{S}}|_1 + \Big(\frac{\mu}{c_\mu} + \eta\beta\, s_L^{\beta-1}\ell_b + \mu\Big)|[\delta\theta_{yx}]_{\bar{S}}|_1 \\ &\quad + \Big(-\frac{\lambda}{c_\lambda} - \eta\beta\, s_L^{\beta-1}\ell_a - \lambda\Big)|[\delta\theta_y]_S|_1 + \Big(\frac{\mu}{c_\mu} + \eta\beta\, s_L^{\beta-1}\ell_b - \mu\Big)|[\delta\theta_{yx}]_S|_1. \end{aligned}$$

Let's pose

$$\underline{c} = \min\left\{\frac{(c_\lambda - 1)\lambda}{c_\lambda} - \eta\beta\, s_L^{\beta-1}\ell_a, \frac{(c_\mu - 1)\mu}{c_\mu} - \eta\beta\, s_L^{\beta-1}\ell_b\right\}$$

and

$$\overline{c} = \max\left\{\frac{(c_\lambda + 1)\lambda}{c_\lambda} + \eta\beta\, s_L^{\beta-1}\ell_a, \frac{(c_\mu + 1)\mu}{c_\mu} + \eta\beta\, s_L^{\beta-1}\ell_b\right\}.$$

We obtain

$$\Delta_n(\widehat{\theta}, \theta^*) \geqslant \underline{c}\,\big(|[\delta\vartheta_y]_{\bar{S}}|_1 + |[\delta\vartheta_{yx}]_{\bar{S}}|_1\big) - \overline{c}\,\big(|[\delta\vartheta_y]_S|_1 + |[\delta\vartheta_{yx}]_S|_1\big). \tag{23}$$

Thus, provided that $\underline{c} > 0$, which is stated in the configuration of the theorem, it only remains to note that, necessarily,

$$\Delta_n(\widehat{\theta}, \theta^*) \leqslant 0$$

since $\widehat{\theta}$ is the global minimizer of $\theta \mapsto R_n(\theta) + \lambda\,|\Omega_y|_1^- + \mu\,|\Delta|_1$. The identification of α given in (12) easily follows. □

Lemma 2.4.11. *Under* ($H_{2.1}$) *and* ($H_{2.2}$)*, the second-order error term of* (19) *satisfies, for* $t = 1$ *and all* $\theta \in N_{r,\alpha}(\theta^*)$*,*

$$e_1(\theta, \theta^*) > \gamma_{r,\eta,\beta,p}\,\|\delta\theta\|_F^2$$

where $\gamma_{r,\eta,\beta,p} > 0$ *is given in* (15)*.*

Proof. From the definition of ϕ in (20) and the fact that $\phi'(0) = \langle\!\langle \nabla R_n(\theta^*), \theta - \theta^* \rangle\!\rangle$, by Taylor-Lagrange

there exists $h \in]0,1[$ satisfying

$$e_1(\theta, \theta^*) = \frac{1}{2}\,\phi''(h). \tag{24}$$

To simplify the calculations, let

$$u_L = \langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\Delta \rangle\!\rangle. \tag{25}$$

We are going to study the behavior of $R_n(\Omega_y, \Delta)$ in the directions $\Omega_y = \Omega_y^* + t\,\delta\theta_y$ and $\Delta = \Delta^* + t\,\delta\theta_{yx}$ through $\phi(t)$, where we recall that $\delta\theta_y = \Omega_y - \Omega_y^*$ and $\delta\theta_{yx} = \Delta - \Delta^*$. One can see that $\phi(t)$ moves from $R_n(\Omega_y, \Delta)$ to $R_n(\Omega_y^*, \Delta^*)$ as t decreases from 1 to 0. The first derivative is

$$\begin{aligned}
\phi'(t) &= \frac{\partial\, R_n(\Omega_y, \Delta)}{\partial t} \\
&= -\langle\!\langle \Omega_y^{-1}, \delta\theta_y \rangle\!\rangle + \langle\!\langle S_y, \delta\theta_y \rangle\!\rangle + 2\langle\!\langle S_{yx}, \delta\theta_{yx} \rangle\!\rangle \\
&\quad + 2\langle\!\langle S_x, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_{yx} \rangle\!\rangle - \langle\!\langle S_x, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta \rangle\!\rangle \\
&\quad + \eta\beta\, u_L^{\beta-1}\big[2\langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_{yx} \rangle\!\rangle - \langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta \rangle\!\rangle\big].
\end{aligned} \tag{26}$$

The second derivative is tedious to write but straightforward to establish,

$$\begin{aligned}
\phi''(t) &= \langle\!\langle \Omega_y^{-1}, \delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y \rangle\!\rangle + 2\big[\langle\!\langle S_x, \delta\theta_{yx}^t\,\Omega_y^{-1}\,\delta\theta_{yx} \rangle\!\rangle - 2\langle\!\langle S_x, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_{yx} \rangle\!\rangle \\
&\quad + \langle\!\langle S_x, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta \rangle\!\rangle\big] \\
&\quad + 2\,\eta\beta\, u_L^{\beta-1}\big[\langle\!\langle L, \delta\theta_{yx}^t\,\Omega_y^{-1}\,\delta\theta_{yx} \rangle\!\rangle - 2\langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_{yx} \rangle\!\rangle \\
&\quad + \langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta \rangle\!\rangle\big] \\
&\quad + \eta\beta(\beta-1)\, u_L^{\beta-2}\big[2\langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_{yx} \rangle\!\rangle - \langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta \rangle\!\rangle\big]^2.
\end{aligned} \tag{27}$$

First, from the combination of Lemmas 2.4.1 and 2.4.9, we clearly have $u_L \geqslant 0$. We also note that $0 \leqslant \|\frac{2}{c}M_1 - cM_2\|_F^2 = \frac{4}{c^2}\,\|M_1\|_F^2 - 4\langle\!\langle M_1, M_2 \rangle\!\rangle + c^2\,\|M_2\|_F^2$ for any $c \neq 0$ and any matrices M_1 and M_2 of same dimensions. It follows, after some reorganizations, that for any $c \neq 0$ and $d \neq 0$,

$$\begin{aligned}
\phi''(t) &\geqslant \langle\!\langle \Omega_y^{-1}, \delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y \rangle\!\rangle + 2\big[\langle\!\langle S_x, \delta\theta_{yx}^t\,\Omega_y^{-1}\,\delta\theta_{yx} \rangle\!\rangle + \langle\!\langle S_x, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta \rangle\!\rangle\big] \\
&\quad - 4\langle\!\langle S_x^{1/2}\,\delta\theta_{yx}^t\,\Omega_y^{-1/2}, S_x^{1/2}\,\Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1/2} \rangle\!\rangle \\
&\quad + 2\,\eta\beta\, u_L^{\beta-1}\big[\langle\!\langle L, \delta\theta_{yx}^t\,\Omega_y^{-1}\,\delta\theta_{yx} \rangle\!\rangle + \langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta \rangle\!\rangle\big] \\
&\quad - 4\langle\!\langle L^{\,1/2}\,\delta\theta_{yx}^t\,\Omega_y^{-1/2}, L^{\,1/2}\,\Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1/2} \rangle\!\rangle \\
&\geqslant \langle\!\langle \Omega_y^{-1}, \delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y \rangle\!\rangle + c_1\langle\!\langle \Omega_y^{-1}, \delta\theta_{yx}\,S_x\,\delta\theta_{yx}^t \rangle\!\rangle + c_2\langle\!\langle S_x, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta \rangle\!\rangle \\
&\quad + \eta\beta\, u_L^{\beta-1}\big[d_1\langle\!\langle \Omega_y^{-1}, \delta\theta_{yx}\,L\,\delta\theta_{yx}^t \rangle\!\rangle + d_2\langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta \rangle\!\rangle\big],
\end{aligned}$$

where $c_1 = 2 - \frac{4}{c^2}$, $c_2 = 2 - c^2$, $d_1 = 2 - \frac{4}{d^2}$ and $d_2 = 2 - d^2$. Here we exploited the previous inequality twice, $u_L \geqslant 0$ and $\beta \geqslant 1$. Moreover, note that from the Lemma 2.4.1 $\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y \in \mathbb{S}_+^q$, and exploiting

also the Lemma 2.4.2 we have that $\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y \in \mathbb{S}_+^q$. From Lemmas 2.4.3, and 2.4.9, we obtain

$$\begin{aligned}\langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\Delta\rangle\!\rangle &= \langle\!\langle \Omega_y^{-1}, \Delta\,L\,\Delta^{\,t}\,\Omega_y^{-1}\,\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\rangle\!\rangle \\ &\leqslant \lambda_{\max}(\Omega_y^{-1}\,\Delta\,L\,\Delta^{\,t})\,\langle\!\langle \Omega_y^{-1}, \delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\rangle\!\rangle \\ &\leqslant \overline{\omega}_L\,\langle\!\langle \Omega_y^{-1}, \delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\rangle\!\rangle,\end{aligned}$$

where $\overline{\omega}_L$ is defined in (9). Replacing L by S_x and $\overline{\omega}_L$ by $\overline{\omega}_S$, a similar bound obviously holds using Lemma 2.4.8. Suppose that c and d are chosen so that $c_1 > 0$, $d_1 > 0$, $c_2 < 0$ and $d_2 < 0$. Then, from Lemma 2.4.9 we have,

$$\begin{aligned}\phi''(t) &\geqslant \langle\!\langle \Omega_y^{-1}, \delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\rangle\!\rangle\,\big[1 - |c_2|\,\overline{\omega}_S - \eta\beta\,u_L^{\beta-1}|d_2|\,\overline{\omega}_L\big] \\ &\qquad + c_1\,\langle\!\langle \Omega_y^{-1}, \delta\theta_{yx}\,S_x\,\delta\theta_{yx}^t\rangle\!\rangle + \eta\beta\,u_L^{\beta-1}d_1\,\langle\!\langle \Omega_y^{-1}, \delta\theta_{yx}\,L\,\delta\theta_{yx}^t\rangle\!\rangle \\ &\geqslant \langle\!\langle \Omega_y^{-1}, \delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\rangle\!\rangle\,\big[1 - |c_2|\,\overline{\omega}_S - \eta\beta\,(p\,\overline{\omega}_L)^{\beta-1}|d_2|\,\overline{\omega}_L\big] \\ &\qquad + c_1\,\langle\!\langle \Omega_y^{-1}, \delta\theta_{yx}\,S_x\,\delta\theta_{yx}^t\rangle\!\rangle + \eta\beta\,(p\,\underline{\omega}_L)^{\beta-1}d_1\,\langle\!\langle \Omega_y^{-1}, \delta\theta_{yx}\,L\,\delta\theta_{yx}^t\rangle\!\rangle.\end{aligned}$$

Now choose $\epsilon_S > 0$ and $\epsilon_L > 0$ small enough so that $\epsilon_S\,\overline{\omega}_S + \eta\beta\,p^{\,\beta-1}\,\overline{\omega}_L^{\,\beta}\,\epsilon_L < 1$ and fix $c = \sqrt{2+\epsilon_S}$ and $d = \sqrt{2+\epsilon_L}$. We finally obtain

$$\phi''(t) \geqslant a_1\,\langle\!\langle \Omega_y^{-1}, \delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\rangle\!\rangle + a_2\,\langle\!\langle \Omega_y^{-1}, \delta\theta_{yx}\,S_x\,\delta\theta_{yx}^t\rangle\!\rangle + a_3\,\langle\!\langle \Omega_y^{-1}, \delta\theta_{yx}\,L\,\delta\theta_{yx}^t\rangle\!\rangle \tag{28}$$

where these positive constants are respectively given by

$$a_1 = 1 - \epsilon_S\,\overline{\omega}_S - \eta\beta\,p^{\,\beta-1}\,\overline{\omega}_L^{\,\beta}\,\epsilon_L, \qquad a_2 = \frac{2\,\epsilon_S}{2+\epsilon_S} \qquad \text{and} \qquad a_3 = \eta\beta\,(p\,\underline{\omega}_L)^{\beta-1}\,\frac{2\,\epsilon_L}{2+\epsilon_L}.$$

The combination of Lemmas 2.4.1, 2.4.3 and 2.4.5 gives, uniformly in $t \in [0,1]$,

$$\langle\!\langle \Omega_y^{-1}, \delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y\rangle\!\rangle \geqslant \lambda_{\min}(\Omega_y^{-1})\,\mathrm{tr}(\delta\theta_y\,\Omega_y^{-1}\,\delta\theta_y) \geqslant \frac{\|\delta\theta_y\|_F^2}{4\,\lambda_{\max}^2(\Omega_y^*)}$$

where the inequality in the denominator comes from $\lambda_{\max}(\Omega_y) \leqslant 2\,\lambda_{\max}(\Omega_y^*)$ already established in the proof of Lemma 2.4.9. Similarly,

$$\langle\!\langle \Omega_y^{-1}, \delta\theta_{yx}\,L\,\delta\theta_{yx}^t\rangle\!\rangle \geqslant \lambda_{\min}(\Omega_y^{-1})\,\mathrm{tr}(\delta\theta_{yx}\,L\,\delta\theta_{yx}^t) \geqslant \frac{\lambda_{\min}(L)\,\|\delta\theta_{yx}\|_F^2}{2\,\lambda_{\max}(\Omega_y^*)}.$$

Lemma 2.4.8 directly enables to bound the last term,

$$\langle\!\langle \Omega_y^{-1}, \delta\theta_y\,S_x\,\delta\theta_y\rangle\!\rangle \geqslant \lambda_{\min}(\Omega_y^{-1})\,\mathrm{tr}(\delta\theta_{yx}\,S_x\,\delta\theta_{yx}^t) \geqslant \frac{\lambda_{\min}(\Sigma_x^*)\,\|\delta\theta_{yx}\|_F^2}{20\,\lambda_{\max}(\Omega_y^*)}.$$

In conclusion, combining (24), (28) and the upper bounds above,

$$\begin{aligned} e_1(\theta,\theta^*) &\geqslant \frac{a_1\,\|\delta\theta_y\|_F^2}{8\,\lambda_{\max}^2(\Omega_y^*)} + \frac{a_2\,\lambda_{\min}(L)\,\|\delta\theta_{yx}\|_F^2}{4\,\lambda_{\max}(\Omega_y^*)} + \frac{a_3\,\lambda_{\min}(\Sigma_x^*)\,\|\delta\theta_{yx}\|_F^2}{40\,\lambda_{\max}(\Omega_y^*)} \\ &\geqslant \min\left\{\frac{a_1}{8\,\lambda_{\max}^2(\Omega_y^*)}, \frac{a_2\,\lambda_{\min}(L)}{4\,\lambda_{\max}(\Omega_y^*)} + \frac{a_3\,\lambda_{\min}(\Sigma_x^*)}{40\,\lambda_{\max}(\Omega_y^*)}\right\}\|\delta\theta\|_F^2 \end{aligned}$$

and we clearly identify $\gamma_{r,\eta,\beta,p} > 0$. □

Lemma 2.4.12. *Assume that λ, μ and η are chosen according to the configuration of the theorem. Suppose also that h_a in* ($\mathrm{C}_{2.1}$) *and h_b in* ($\mathrm{C}_{2.2}$) *satisfy*

$$\max\{h_a, h_b\} < \frac{r^*\,\gamma_{r,\eta,\beta,p}}{c_{\lambda,\mu}\sqrt{|S|}}$$

where r^* is given in (14), *$\gamma_{r,\eta,\beta,p}$ in* (15) *and $c_{\lambda,\mu}$ in* (16). *Then, under* ($\mathrm{H}_{2.1}$) *and* ($\mathrm{H}_{2.2}$), *the estimation error satisfies $\|\delta\vartheta\|_F \leqslant r^*$.*

Proof. By convexity of the objective and optimality of $\widehat{\theta}$, each move from θ^* in the direction $t\,\delta\vartheta$ for $t \in [0,1]$ must lead to a decrease of the objective, *i.e.*

$$R_n(\theta^* + t\,\delta\vartheta) + \lambda\,|\Omega_y^* + t\,\delta\vartheta_y|_1^- + \mu\,|\Delta^* + t\,\delta\vartheta_{yx}|_1 - R_n(\theta^*) - \lambda\,|\Omega_y^*|_1^- - \mu\,|\Delta^*|_1 \leqslant 0.$$

Taking the notation of (23), this can be rewritten as $\Delta_n(\theta^* + t\,\delta\vartheta, \theta^*) \leqslant 0$. If $\|\delta\vartheta\|_F \leqslant r^*$ then choose $t = 1$, otherwise calibrate $0 < t < 1$ such that $\|t\,\delta\vartheta\|_F = r^*$. Then, from Lemma 2.4.10, it clearly follows that $\theta^* + t\,\delta\vartheta \in N_{r,\alpha}(\theta^*)$. Hence, the reasoning preceding (23) still holds. Taking up the said reasoning, and exploiting the fact that$R_n(\theta^* + t\,\delta\vartheta) - R_n(\theta^*) = t\,\phi'(0) + e_t(\theta,\theta^*)$, we can see that by Taylor-Lagrange, there exists $h \in\,]0,1[$ such that

$$\begin{aligned} \Delta_n(\theta^* + t\,\delta\vartheta, \theta^*) &= t\,\phi'(0) + \lambda\,|\Omega_y^* + t\,\delta\vartheta_y|_1^- + \mu\,|\Delta^* + t\,\delta\vartheta_{yx}|_1 - \lambda\,|\Omega_y^*|_1^- - \mu\,|\Delta^*|_1 + e_t(\theta,\theta^*) \\ &\geqslant \underline{c}\left(|[t\,\delta\vartheta_y]_{\bar{S}}|_1 + |[t\,\delta\vartheta_{yx}]_{\bar{S}}|_1\right) - \overline{c}\left(|[t\,\delta\vartheta_y]_S|_1 + |[t\,\delta\vartheta_{yx}]_S|_1\right) + \frac{t^2}{2}\phi''(h). \end{aligned}$$

From Lemma 2.4.11, we obtain

$$\begin{aligned} 0 &\geqslant \underline{c}\left(|[t\,\delta\vartheta_y]_{\bar{S}}|_1 + |[t\,\delta\vartheta_{yx}]_{\bar{S}}|_1\right) - \overline{c}\left(|[t\,\delta\vartheta_y]_S|_1 + |[t\,\delta\vartheta_{yx}]_S|_1\right) + \gamma_{r,\eta,\beta,p}\,\|t\,\delta\vartheta\|_F^2 \\ &\geqslant -\overline{c}\,|[t\,\delta\vartheta]_S|_1 + \gamma_{r,\eta,\beta,p}\,\|t\,\delta\vartheta\|_F^2. \end{aligned}$$

where we used $\underline{c} > 0$. Note that the configuration of the penalization parameters, which is stated in the theorem, implies the following relations

$$\frac{(c_\lambda+1)d_\lambda\,h_a}{c_\lambda} \geqslant \frac{(c_\lambda+1)\lambda}{c_\lambda}, \qquad \frac{(c_\mu+1)d_\mu\,h_b}{c_\mu} \geqslant \frac{(c_\mu+1)\mu}{c_\mu}, \qquad e_\lambda\,h_a \geqslant \eta\,\beta\,s_L^\beta\,\ell_a \quad \text{and} \quad e_\mu\,h_b \geqslant \eta\,\beta\,s_L^\beta\,\ell_b.$$

From Cauchy-Schwarz inequality $|[\cdot]_S|_1^2 \leqslant |S|\,\|[\cdot]_S\|_F^2$, we finally obtain

$$0 \quad \geqslant \quad -c_{\lambda,\mu}\max\{h_a,h_b\}\sqrt{|S|}\,\|t\,\delta\vartheta\|_F + \gamma_{r,\eta,\beta,p}\,\|t\,\delta\vartheta\|_F^2$$

where constant $c_{\lambda,\mu}$ is given in (16). Note that in the proof of Lemma 2.4.10, it was sufficient to see that $R_n(\theta) - R_n(\theta^*) \geqslant \phi'(0)$ whereas here, we must consider $R_n(\theta) - R_n(\theta^*) = \phi'(0) + e_1(\theta,\theta^*)$ to meet our purposes. That explains the presence of $\gamma_{r,\eta,\beta,p}\,\|t\,\delta\vartheta\|_F^2$ in the inequality. We deduce that the error must satisfy

$$\|t\,\delta\vartheta\|_F \leqslant \frac{c_{\lambda,\mu}\sqrt{|S|}\max\{h_a,h_b\}}{\gamma_{r,\eta,\beta,p}}.$$

As a corollary, it holds that $\|\delta\vartheta\|_F > r^* \;\Rightarrow\; c_{\lambda,\mu}\sqrt{|S|}\max\{h_a,h_b\} \geqslant r^*\,\gamma_{r,\eta,\beta,p}$ or, conversely written, $c_{\lambda,\mu}\sqrt{|S|}\max\{h_a,h_b\} < r^*\,\gamma_{r,\eta,\beta,p} \;\Rightarrow\; \|\delta\vartheta\|_F \leqslant r^*$. □

Lemma 2.4.13. *Assume that λ, μ and η are chosen according to the configuration of the theorem. Then, under* ($\mathrm{H}_{2.1}$)*, there exists absolute constants $b_1 > 0$ and $b_2 > 0$ such that, for any $b_3 \in]0,1[$ and as soon as*

$$n \geqslant \max\left\{b_1\,(q + \lceil s_\alpha\rceil \ln(p+q)),\, \ln(10(p+q)^2) - \ln(b_3)\right\},$$

with probability no less that $1 - \mathrm{e}^{-b_2 n} - b_3$ both the random hypothesis ($\mathrm{H}_{2.2}$) *is satisfied and the upper bound*

$$\max\{h_a,h_b\} \leqslant 16\,m^*\sqrt{\frac{\ln(10(p+q)^2) - \ln(b_3)}{n}}$$

holds, where h_a and h_b are given in ($\mathrm{C}_{2.1}$) *and* ($\mathrm{C}_{2.2}$)*, s_α is defined in* (13) *and m^* in* (17)*. Hence, one can find a minimal number of observations n_0 such that the theorem holds with high probability as soon as $n > n_0$.*

Proof. All the ingredients of the proof are established in [YZ14]. The authors start by recalling that there exists absolute constants $b_1 > 0$ and $b_2 > 0$ such that hypothesis ($\mathrm{H}_{2.2}$) is satisfied with probability no less than $1 - \mathrm{e}^{-b_2 n}$ as soon as $n \geqslant b_1\,(q + \lceil s_\alpha\rceil \ln(p+q))$. We also refer the reader to Lem. 5.1 and Thm. 5.2 of [BDDW08], or to Lem. 7.4 of [Gir14] for the random bounds of the restricted isometry constants. Afterwards, they prove (see Prop. 4) that, as soon as $n \geqslant \ln(10(p+q)^2) - \ln(b_3)$ for some $b_3 > 0$, with probability $1 - b_3$,

$$\max\{h_a,h_b\} \leqslant 16\,m^*\sqrt{\frac{\ln(10(p+q)^2) - \ln(b_3)}{n}}.$$

To find the minimal number of observations, we just need to make sure that the above bound is itself smaller than the one of Lemma 2.4.12. It is then not hard to see that we may retain the minimal size n_0 given in (6). □

2.5 Simulations and real dataset

Our estimation algorithm consists of a coordinate descent procedure. We exploit the fact that the minimization problem (5) is jointly convex, and alternate between the computation of

$$\widehat{\Omega}_y = \arg\min_{\mathbb{S}^q_{++}} \ell_{\lambda,\mu,\eta}(\Omega_y, \widehat{\Delta}) \quad \text{and} \quad \widehat{\Delta} = \arg\min_{\mathbb{R}^{q\times p}} \ell_{\lambda,\mu,\eta}(\widehat{\Omega}_y, \Delta).$$

Each step is done by an Orthant-Wise Limited-Memory Quasi-Newton (OWL-QN) algorithm (see *e.g.* [AG07]). The first subproblem is performed through half-vectorization (vech) to ensure symmetry and we set the objective to $+\infty$ on $\bar{\mathbb{S}}^q_{++}$ to ensure positive definiteness of the solution. The coordinate descent is stopped when

$$\|\widehat{\Omega}_y^{(t)} - \widehat{\Omega}_y^{(t-1)}\|_2 \leqslant \epsilon \max(1, \|\widehat{\Omega}_y^{(t-1)}\|_2) \quad \text{and} \quad \|\widehat{\Delta}^{(t)} - \widehat{\Delta}^{(t-1)}\|_2 \leqslant \epsilon \max(1, \|\widehat{\Delta}^{(t-1)}\|_2)$$

following two consecutive iterations $t-1$ and t, where $\epsilon > 0$ is a small threshold depending on the desired precision.

We are now going to try our method on synthetic data first, and then on a real dataset. We will pay attention to the role played by β, in particular we will see that it can be useful as well as counterproductive, depending on the situations.

2.5.1 Simulations

For each scenario, we first generate i.i.d. standard Gaussian vectors $X_i \in \mathbb{R}^p$, then $Y_i \in \mathbb{R}^q$ is simulated according to the setting and we estimate Ω_y and Δ. From the relations detailed in Section 2.1, we recall that $Y_i = B^t X_i + E_i$ with $E_i \sim \mathcal{N}(0, R)$ is an equivalent formulation, provided that $B = -\Delta^t \Omega_y^{-1}$ and $R = \Omega_y^{-1}$. In a compact form, we may also write

$$Y = XB + E \quad \text{or} \quad \operatorname{vec}(Y) = (I_q \otimes X)\operatorname{vec}(B) + \operatorname{vec}(E)$$

where the i-th row of Y is ${Y_i}^t$ and the i-th row of X is X_i^t. Thus, we can estimate B using the Lasso (Las) and the Group-Lasso (GLas) in the vectorized form, to provide a basis for comparison between our method and the usual penalized methods. The Lasso penalty is obviously $\|\operatorname{vec}(B)\|_1$ to promote coordinate sparsity while, for the Group-Lasso, we use the penalty $\|B_1\|_2 + \ldots + \|B_p\|_2$ where B_i is the i-th row of B, to promote row sparsity and exclude altogether some predictors from the model. We also implement some variants of our generalized graphical model (GenGm):

- the case where $\Omega_y = R^{-1}$ is known and does not need to be estimated is the Oracle (Or) ,
- the case where $\eta = 0$ so that β has no influence is the classic PGGM (Gm),
- the case where $\lambda = 0$ and $\beta = 1$ is called the Spring (Spr) by the authors of [CMHR17].

We will focus on structured scenarios. With no structure in Δ, there is no reason why our method should outperform the usual PGGM. In a completely random setting, we have observed that all PGGM procedures perform identically. In fact, a slight gain can be obtained compared to Spr and Gm simply

due to the flexibility induced by the additional parameter (Spr and Gm are particular cases of GenGm). However, that clearly cannot counterbalance the extended computational times, and GenGm should not be used for such situations. The calibration of the regularization parameters is made using a cross-validation on a training set of size $n_t = 150$ and the accuracy is evaluated thanks to the mean squared prediction error (MSPE) on a validation set of size $n_v = 1000$,

$$\text{MSPE} = \frac{\left\| Y + X\,\widehat{\Delta}^{\,t}\,\widehat{\Omega}_y^{-1} \right\|_F^2}{q\,n_v}. \tag{29}$$

Due to the large amount of treatments, the grids for cross-validation are not very sharp here but they will be carefully refined for the real datasets of the next section. The covariance between the outputs is $R = (r^{\,|i-j|})_{1 \leqslant i,\, j \leqslant q}$ for $r = \frac{1}{2}$ and we work with $p = 100$. Each scenario is repeated $N = 500$ times and GenGm is evaluated with numerous values of β, from 0.25 to 2 with a step of 0.25. The results of the following scenarios are summarized on Figures 2.2, 2.3 and 2.4 below, respectively.

→ Scenario 1 ($q = 1$). We draw $\omega_i = \pm\frac{1}{2}$ for $i = 1, \ldots, 10$ and we fill 10 randomly selected sections of size 3 in Δ with ω_i. The remaining part of Δ is 0.

→ Scenario 2 ($q = 2$). We draw $\omega = \pm\frac{1}{2}$ and one randomly selected row of Δ is filled with ω while the other is identically 0.

→ Scenario 3 ($q = 3$). We draw $\omega_i = \pm\frac{1}{2}$ and we fill a randomly selected section of size 30 on the i-th row of Δ with ω_i, for $i = 1, 2, 3$. The remaining part of Δ is 0.

The row structure is promoted by a normalized first finite difference operator

$$L = \frac{1}{2}\begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ -1 & 2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 2 & -1 \\ 0 & \ldots & 0 & -1 & 1 \end{pmatrix} \tag{30}$$

which, through $\Delta\,L\,\Delta^{\,t}$, tends to penalize the difference between two consecutive values on a same row (as does Fused-Lasso with ℓ_1 penalty). Yet, the Fused-Lasso is not a suitable alternative to GLas and Las in this precise context because $B = -\Delta^{\,t}\,\Omega_y^{-1}$ is not supposed to have a row structure even if Δ has one. For this choice of L, one can note that, in the particular case where $R = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$,

$$\langle\!\langle L, \Delta^{\,t}\,\Omega_y^{-1}\,\Delta \rangle\!\rangle^{\beta} = \left(\sum_{i=1}^{q} \sigma_i^2 \sum_{j=2}^{p} (\omega_{i,j} - \omega_{i,j-1})^2 \right)^{\beta} \geqslant \sum_{i=1}^{q} {\sigma_i^2}^{\beta} \sum_{j=2}^{p} \left|\omega_{i,j} - \omega_{i,j-1}\right|^{2\beta}$$

where $\omega_{i,j}$ is the (i,j)-th element of Δ, so we may fairly expect that $\beta \geqslant 1$ is going to strengthen the smoothness of the estimation and to enforce all the more the structuring.

Remark 2.5.1 (Validity of the hypotheses)**.** We could as well add a small diagonal element in the matrix L defined above, positive semi-definite but not invertible. The resulting effect would be a negligible ridge-

like penalization on the elements of Δ. This is not required for the estimation procedure but useful for Theorem 2.3.1 to hold (see *e.g.* ($H_{2.1}$)). Likewise, it seemed interesting to test some settings with $\beta < 1$ even if the theory developped in the chapter does not give any guarantee for them, as a basis for comparison.

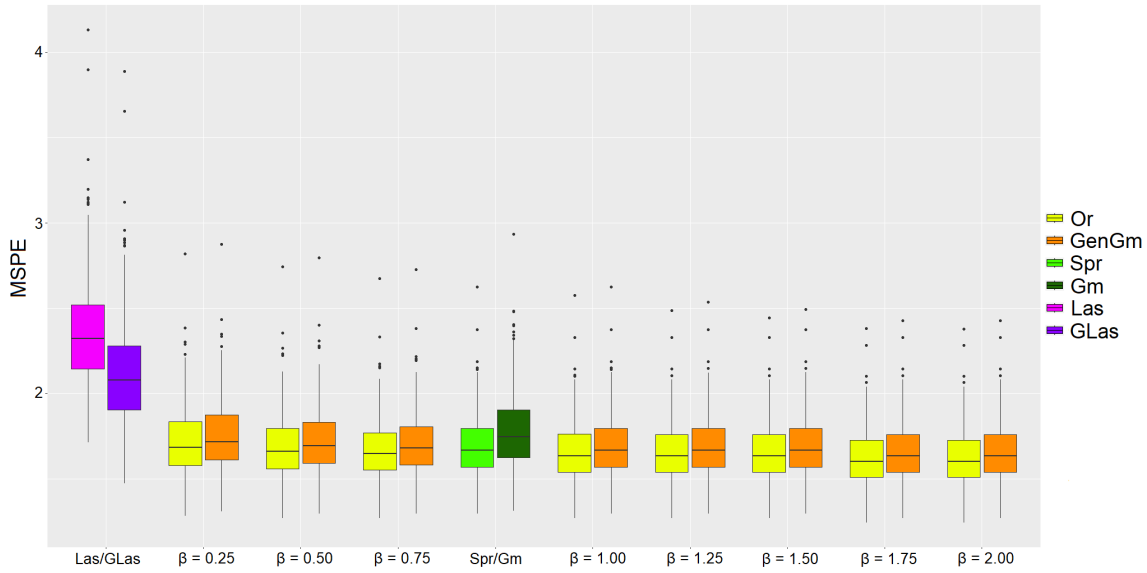


FIGURE 2.2 : Mean squared prediction error for $N = 500$ repetitions of the weakly structured Scenario 1.

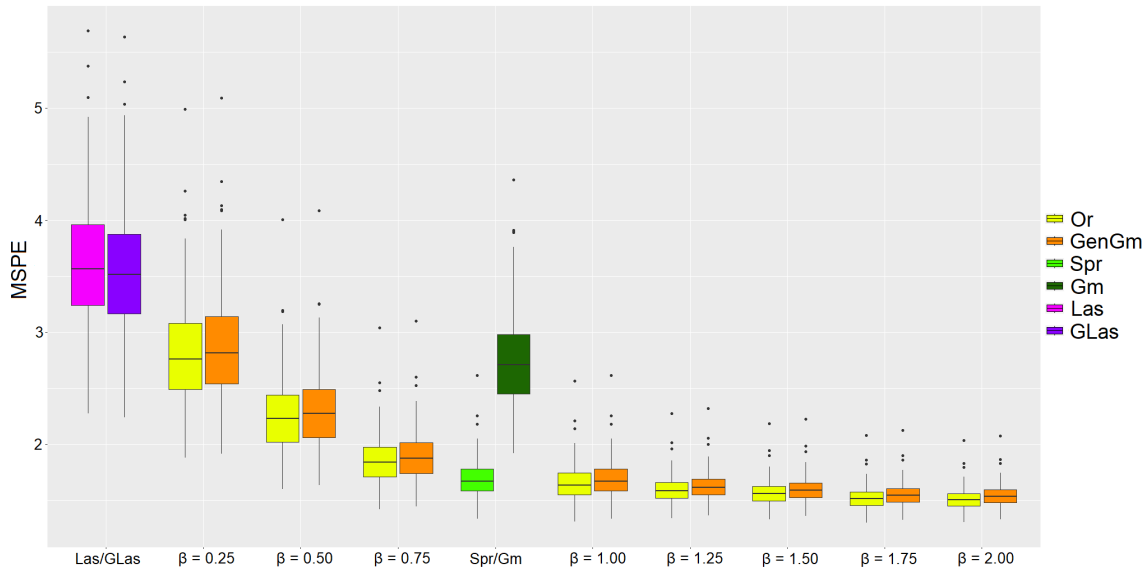


FIGURE 2.3 : Mean squared prediction error for $N = 500$ repetitions of the strongly structured Scenario 2.

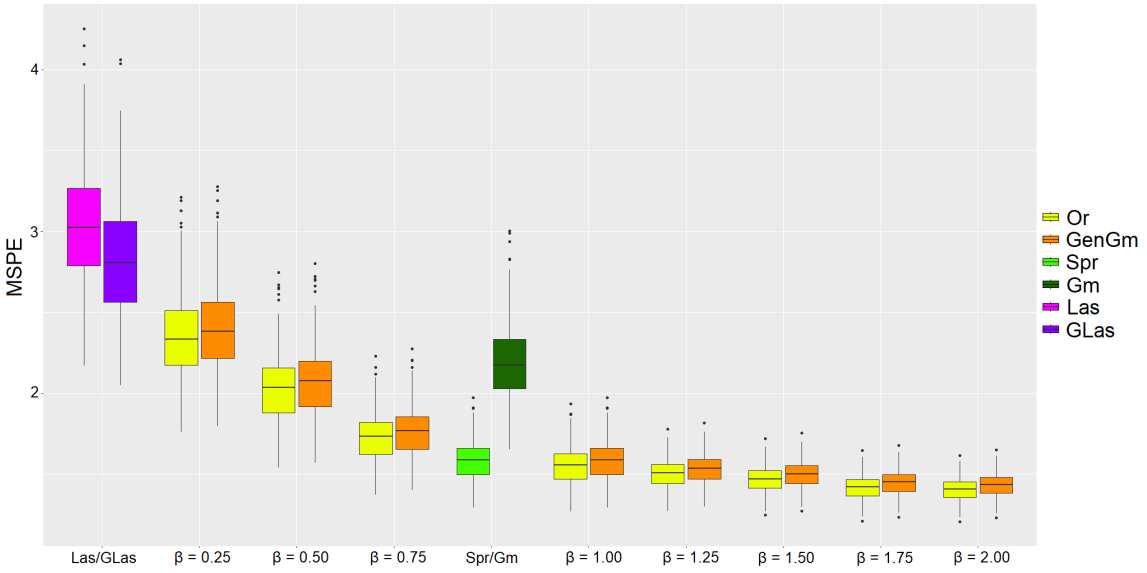


FIGURE 2.4 : Mean squared prediction error for $N = 500$ repetitions of the strongly structured Scenario 3.

First of all, one can observe that Las and GLas are left behind in all our simulations. This is not surprising since the covariance between the outputs cannot be recovered with the standard Lasso, at least for $q \geqslant 2$. Generally, GLas remains more robust compared to Las, probably due to the high level of sparsity in Δ approximately passed to B (provided that the covariances in R are small enough), and exploited by the grouping effect. In the weakly structured setting (Scenario 1), we also observe that, as expected, all PGGM procedures perform almost identically, with obviously an advantage for Or (although small, illustrating the accuracy of the estimation). In the strongly structured settings (Scenarios 2 and 3), Gm gives results below the expected level, because it is not designed to promote such layouts. On the contrary, thanks to this choice of L showing here great efficiency, GenGm and Spr are doing pretty well. Note that, in this context, GenGm with $\beta = 1$ is almost the same as Spr since, q being small, λ does not play a crucial role. However, some empirical facts draw our attention: the prediction error decreases with β to some extent, but the most interesting fact seems to be the simultaneous decrease of its variance. It is likely that the increasing pressure exerted by β on the estimation procedure leads to a higher homogeneity in the numerical results, despite the repetitions of random experiments under random settings. In other words, the structuring seems to be strengthened and we also observe that the convergence of the algorithm is faster, which logically follows from the latter remarks (especially clear when we compare $\beta = 0.25$ and $\beta = 2$). On the other hand, for the opposite reason, we notice that the predictions are hardly better than Gm (even worse in some cases), both on average and in terms of variability, for $\beta < 1$, and these simulations tend to undermine such values of the hyperparameter. On the whole, GenGm with $\beta > 1$ might be a sound approach for practitioners who place a high priority on structuring the estimations, even if Remark 2.5.2 below should probably temper this statement. To conclude, let us consider the strongly structured scenarios with $L = I_p$ (without structuring) in the Oracle setting with $\beta = 2$, and let us compare the results with those of Figures 2.3 and 2.4, obtained with the correct version of L given in (30). The results are displayed on Figure 2.5 where we can see that the

benefit of structuring is manifest. Unsurprisingly, the results without structuring are close to those of Gm since $L = I_p$ only strengthens the shrinkage effect with ridge-like additional penalties.

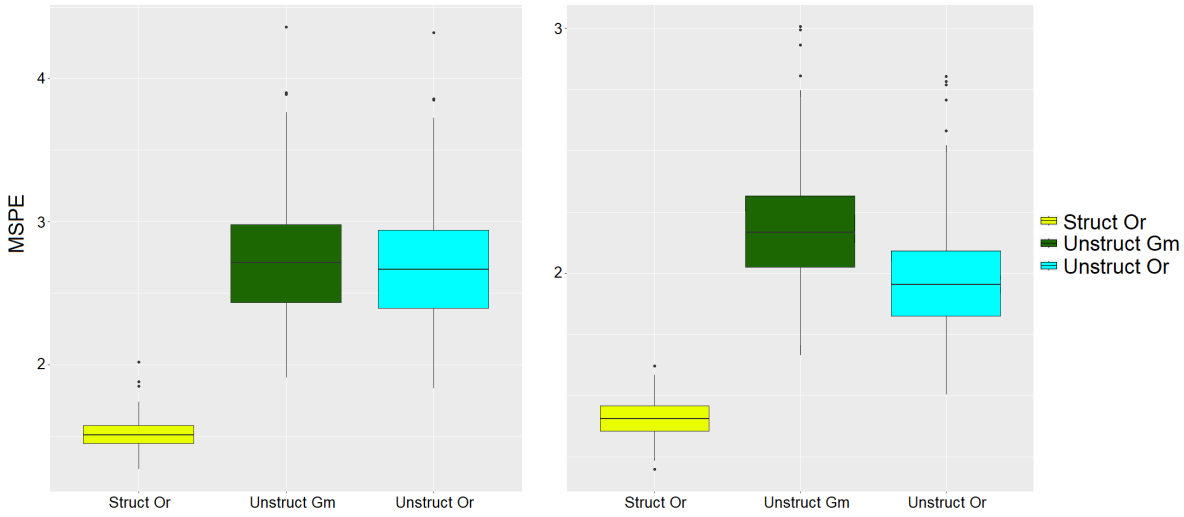


FIGURE 2.5 : Mean squared prediction error for $N = 500$ repetitions of the strongly structured Scenario 2 (left) and Scenario 3 (right) for Or, Gm and the unstructured Or ($L = I_p$), with $\beta = 2$.

Remark 2.5.2 (Computational time)**.** To estimate (Ω_y, Δ) in the model Spr, the authors of [CMHR17] use a very judicious and efficient method relying, in each step of the coordinate descent procedure, on a direct computation of the estimation of Ω_y together with an Elastic-Net estimation of Δ. This is possible for $\lambda = 0$ and $\beta = 1$, but unfortunately cannot be implemented in the general setting. As a result, computational times remain an issue that should be paid attention to.

Remark 2.5.3 (Oracle-type errors)**.** The mean value of the estimation errors $\|\widehat{\Delta} - \Delta\|_F^2$ leads to the same kind of observations for the models being compared in the simulations. But the minimal prediction error does not always coincide with an optimal support recovery due to the shrinkage effect on the estimation of Δ, when the coefficients or the covariates are not very contrasting. The so-called F-score is given by

$$F = \frac{2\, p_r\, r_e}{p_r + r_e} \quad \text{where} \quad p_r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad \text{and} \quad r_e = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

are the *precision* and the *recall*, respectively, and where T/F and P/N stand for true/false and positive/negative. In the strongly structured scenarios, F is generally located between 0.60 and 0.65, and a deeper analysis shows that a proportion of more than 0.99 of true non-zero values are recovered (that is, the part of the true active set S related to Δ). If the models are not calibrated to reach the best prediction error but the best F-score, F regularly exceeds 0.90, at least for the structured procedures.

Nevertheless, Scenarios 2 and 3 are very strongly structured, more than one would expect from an unknown underlying generating process, and the real dataset of the next section is going to highlight the fact that the improvement may be hardly noticeable with respect to β. But we will see that β can still be useful for variable selection.

2.5.2 A real dataset

The dataset available as `CanadianWeather` in the `R` package `fda` contains daily temperature and precipitation at 35 different locations in Canada, averaged over annual reports starting in 1960 and ending in 1994 (see *e.g.* [RS06]). We intend to look at the direct links between the minimal and maximal rainfall (on the $\log_{10}$ scale) and the temperature pattern in the 35 weather stations, so as to identify the times of the year that have a strong effect on rainfall (positive as well as negative). In this context, $n = 35$, $q = 2$ and $p = 365$. Figure 2.6 shows temperature and log-precipitation measured over a year in Montreal, chosen as an example, together with the empirical distribution of the minimal and maximal log-precipitation for the 35 weather stations. We can note that, since the data are averaged over numerous years, outliers are unlikely even for the extremes (min and max).

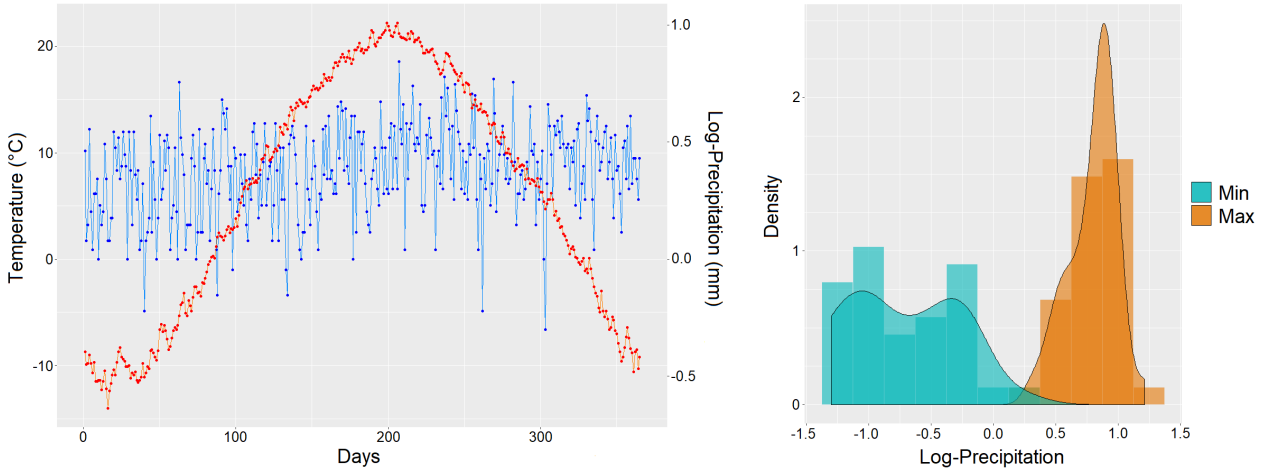


FIGURE 2.6 : Temperature and log-precipitation measured over a year in Montreal (left). Empirical distribution of the minimal and maximal log-precipitation for the 35 weather stations (right).

Some authors (see *e.g.* [Sla12]) have already highlighted the pertinence of using the matrix L defined in (30) in this dataset, because the predictors are ordered temporally so that the selection of isolated days instead of relevant sequences of days seems an unreliable procedure for statistical interpretation. To assess the models, we repeat $N = 100$ times the following experiment:

1. we randomly select $n_t = 25$ observations which constitute the training set on which we first perform a 2-fold cross-validation for parameter calibration, followed by the model estimation ;

2. the remaining $n_v = 10$ observations then constitute the validation set, and are used to compute the MSPE (29) related to the prediction of the minimum ($\min_p$) and maximum ($\max_p$) precipitation.

We can see on Figure 2.7 that all structured PGGM perform almost identically, with the phenomenon described in the previous section still visible but to a lesser extent. We can even notice that structuring is hardly beneficial for this dataset, from a purely numerical point of view. This conclusion can also be found in [Sla12], where the author compares the structured Elastic-Net with unstructured alternatives to predict the 0.25-, 0.50- and 0.75-quantiles of the log-precipitation, through independent regressions.

But we will see that, in terms of variable selection and statistical interpretation, L and β still have a substancial role to play.

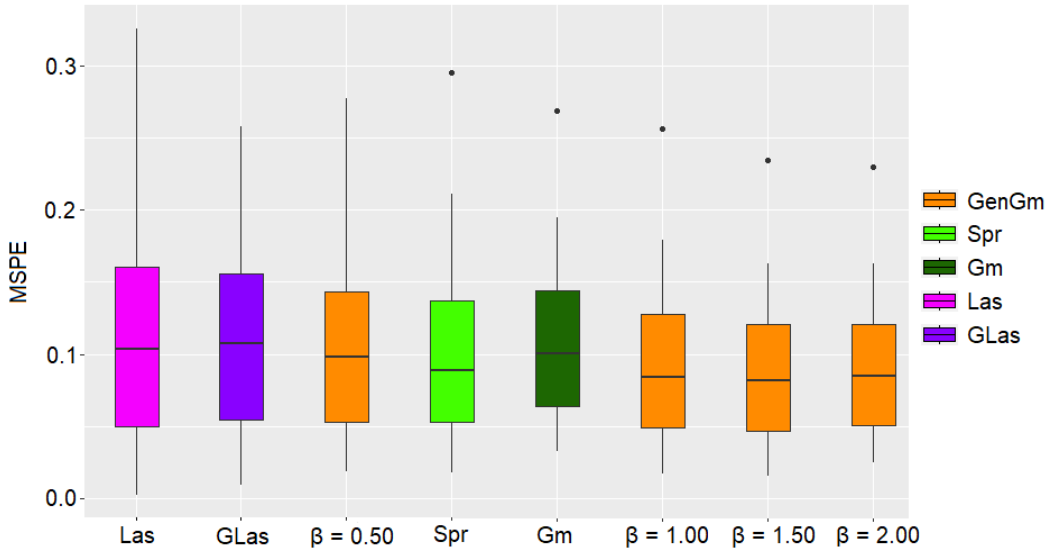


FIGURE 2.7 : Mean squared prediction error for $N = 100$ repetitions of the experiment. GenGm for $\beta \in \{0.5, 1, 1.5, 2\}$ is compared with Spr, Gm, Las and GLas.

The point is that we have observed that the best prediction error does not usually coincide with a sparse solution (see Remark 2.5.3 above) when the coefficients or the covariates are not very contrasting. In particular, this was the case of our simulation study with $\pm\frac{1}{2}$ coefficients and $\mathcal{N}(0,1)$ covariates. So, just as they look at the Lasso's regularization paths, practitioners may choose the desired degree of sparsity, depending on p/n, by playing with the hyperparameters. Here, on the basis of the MSPE, most of the time we must retain $\mu \ll 10^{-2}$ and only a few direct links are set to zero. To look for sequences of days directly related to $\min_{\mathrm{p}}$ and $\max_{\mathrm{p}}$, we decided to constraint $\mu \geqslant 10^{-2}$ and focus on variable selection. The active set of Δ is evaluated on the basis of $n_t = 25$ randomly chosen observations. The experiment is repeated $N = 100$ times, and the locations having a frequency of occurrence that exceeds 0.5 are retained (or, equivalently, those whose estimates have a non-zero median). This can be seen as a measure of variable importance. The results are given on Figures 2.8 and 2.9 below for $\min_{\mathrm{p}}$ and $\max_{\mathrm{p}}$, respectively, with a fixed set of regularization parameters and increasing values of β. The objective is to show the influence of the latter, all other things being equal. The colored areas highlight the days having a frequency of occurrence, represented by gray crosses, that exceeds 0.5 in the $N = 100$ repetitions of the experiment. Note that, since we retain $\lambda = 0$ in these experiments, GenGm for $\beta = 1$ coincides with Spr. We can see that the increasing pressure exerted by β on the estimation procedure tends to refine the selection by giving priority to the most important variables and by dropping the others much more easily, at the cost of prediction results: we are undoubtedly in a selection process. The sequence of inclusions

$$\widehat{S}_{\beta_2} \subset \widehat{S}_{\beta_1} \qquad \text{for } \beta_1 < \beta_2$$

that we observe for the estimated active sets is clearly a guarantee of quality for the selected variables.

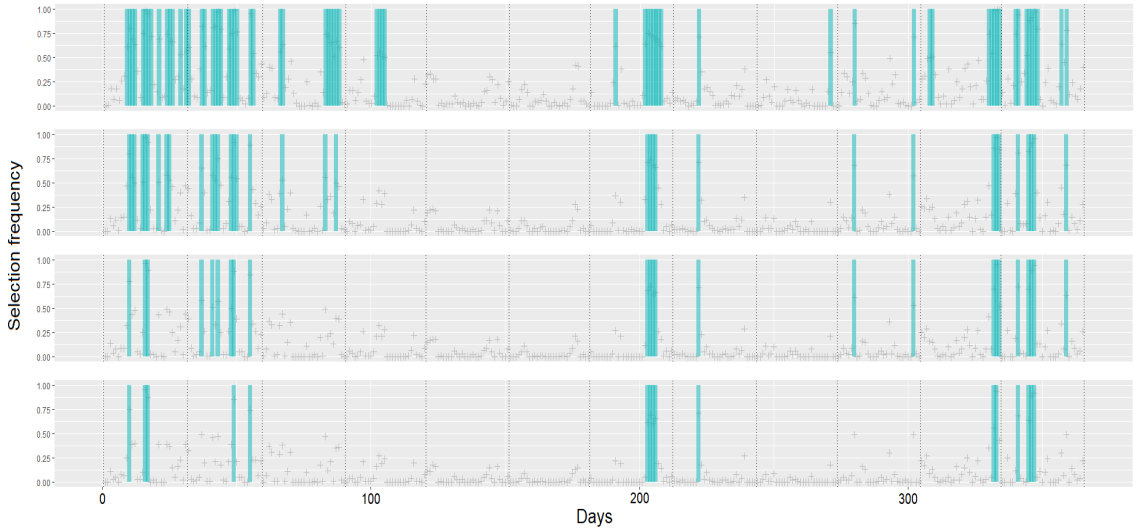


FIGURE 2.8 : Variable selection for $\min_p$ by GenGm with $(\lambda, \mu, \eta) = (0, 0.05, 1)$ and, from top to bottom, $\beta \in \{0.5, 1, 1.5, 2\}$.

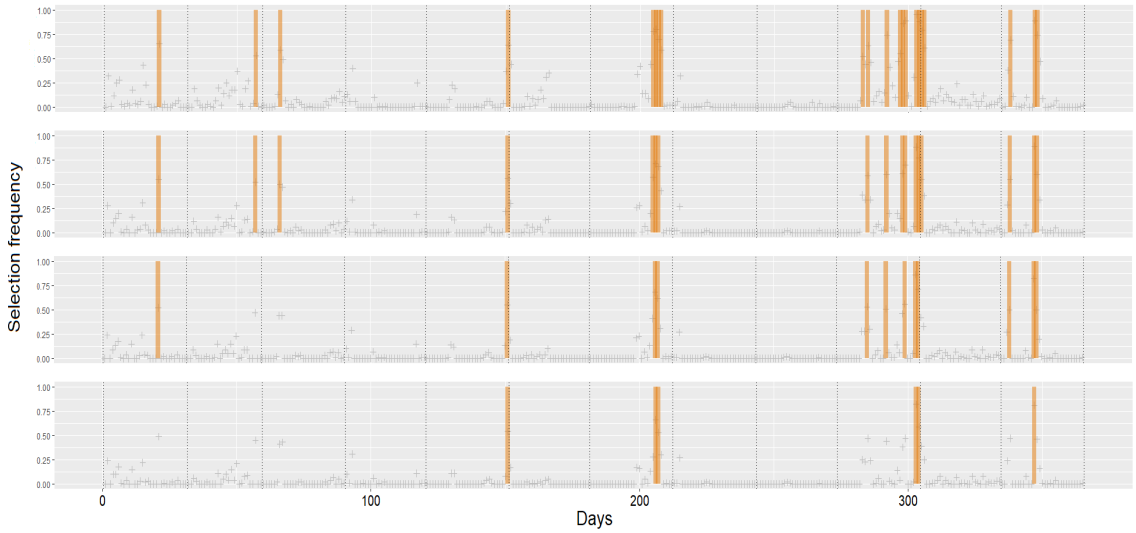


FIGURE 2.9 : Variable selection for $\max_p$ by GenGm with $(\lambda, \mu, \eta) = (0, 0.05, 1)$ and, from top to bottom, $\beta \in \{0.5, 1, 1.5, 2\}$.

The median values of the estimated direct links between the temperature of the days and the pair $(\min_p, \max_p)$ are represented on Figure 2.10 together with the estimated regression coefficients, for $\beta = 2$. We recall that the relation $B = -\Delta^t \Omega_y^{-1}$ simply lead to

$$\widehat{B} = -\widehat{\Delta}^t \widehat{\Omega}_y^{-1}.$$

We detect sequences of influent days in November, December, January and February, especially related to $\min_p$, positively at the end of the year and negatively at the beggining. This is broadly consistent with

the analysis of [Sla12] – even if the responses are not extremes but quantiles in it – with however two differences: the regression coefficients associated with $\max_p$ are much lower compared to $\min_p$ whereas it is not that clear in the reference, and an activity is also detected between July and August. The main explanation, at least for the first of them, probably lies in the use of graphical models that take into account the correlation between responses. Indeed, as can be seen on Figure 2.11 which gives an overview of the estimation of R obtained from the repeated experiments, a non-zero correlation is detected between the responses (≈ 0.32). The influence of November and December on all quantiles and that of January and February on the 0.75-quantile in [Sla12] might actually be an artificial effect of the correlation with the 0.25-quantile. This is what our study suggests by highlighting $\min_p$ compared to $\max_p$: the 'real' effect appears to be on $\min_p$ whereas $\max_p$ seems to react only through a phenomenon of correlation with $\min_p$. From this point of view, the interest of graphical models instead of independent regressions is particularly obvious.

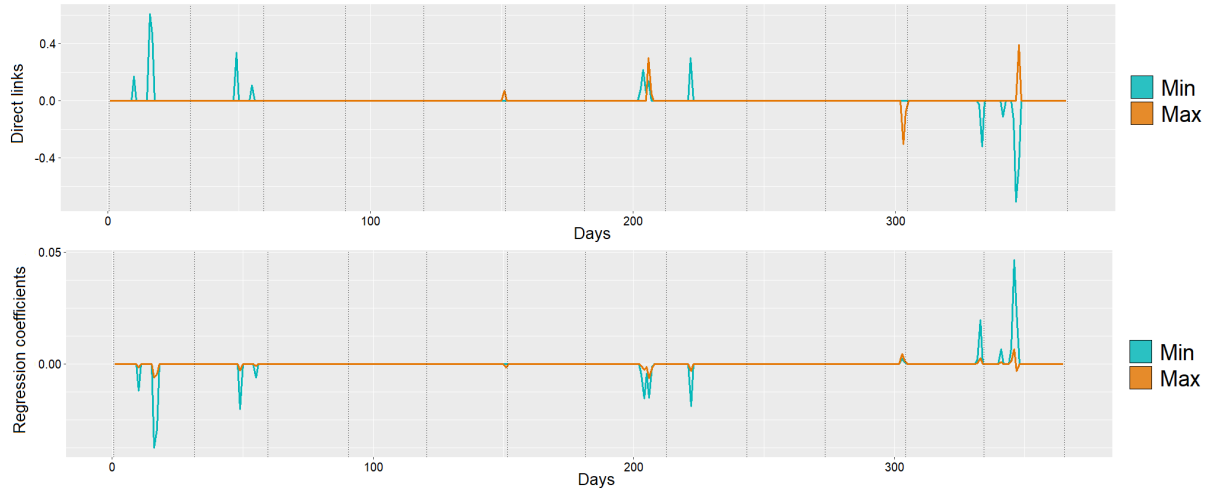


FIGURE 2.10 : Estimated direct links (top) and regression coefficients (bottom) for the pair $(\min_p, \max_p)$ by GenGm with $(\lambda, \mu, \eta) = (0, 0.05, 1)$ and $\beta = 2$, after the $N = 100$ experiments. Dotted lines divide the panel into months.

Let us also mention that, interestingly enough, we notice that the role of η tends to depreciate for the large values of β. For example, for the same regularization parameters $(\lambda, \mu) = (0, 0.05)$ and $\beta = 2$, the difference between the estimated active sets for $\eta = 0.1$ and $\eta = 1$ is almost negligible (depending on the experiments, between 1 and 3 days are concerned, on average). Based on these studies and observations, we might conclude that β is insignificant when we are interested in the best prediction error on a validation set (even counterproductive with respect to computational times, *e.g.* compared to Spr), whereas it seems to have a substancial role to play when focusing on selection, by accelerating the discrimination of variables. In the first case, η has to be carefully adjusted while in the second case, β will quickly help to reach the desired sparsity.

Remark 2.5.4 (Structure matrix)**.** For the simulations and the real dataset, we have used the popular first finite difference operator given in (30). Other examples can be found in the literature, like the promotion of a genetic distance for genomic selection in *Brassica napus* [CMHR17] or the bidimensional

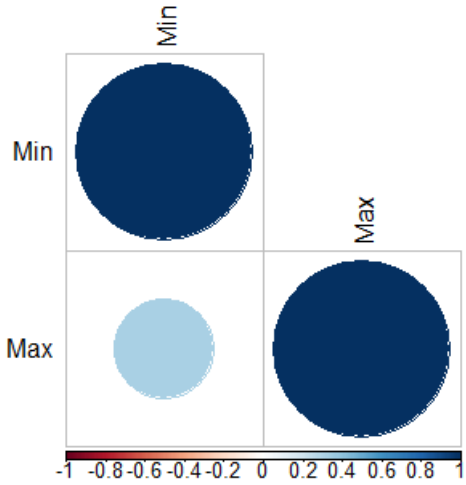


FIGURE 2.11 : Estimated correlation between $\min_p$ and $\max_p$ by GenGm with $(\lambda, \mu, \eta) = (0, 0.05, 1)$ and $\beta = 2$, after the $N = 100$ experiments. The off-diagonal entry is approximately 0.32.

discretization of the Laplacian to work on handwritten digit recognition [Sla12]. More generally, L can be used in a classic Bayesian prior supposed to promote some covariance structure on the direct links, with no 'physical' structuring in mind (like temporal, spatial or genetic proximity).

2.6 Conclusion

In conclusion, our work is a generalization of [YZ14], using the same technical tools to establish an upper bound on the estimation error when a prior on the direct links generates an additional structural penalty in the objective, provided that the model is suitably regularized. Our work is also an improvement of [CMHR17] since, while being inspired by the methodology of the authors, we generalize the prior and give some theoretical guarantees. The empirical study shows that the hyperparametrization in the prior, although more expensive in adjusting the parameters, is likely to refine the selection results but clearly, this does not appear as a crucial improvement compared to the two previous points. Let us conclude the chapter by highlighting two weaknesses that might be trails for future studies. On the one hand, the Laplace distribution is often used as a prior in the Bayesian Lasso (see *e.g.* Sec. 6.1 of [HTW15]). However, our reasonings do not allow $\beta = 1/2$, which may correspond to a multivariate Laplace distribution on the direct links. Combined with the first finite difference operator L, the choice $\beta = 1/2$ could generate a Fused-Lasso-type penalty. In this regard, it would be challenging and interesting to obtain some theoretical guarantees for $\beta \geqslant 1/2$ and not only for $\beta \geqslant 1$, even if our probably too brief simulation study does not encourage the choice of $\beta < 1$. On the other hand, $\lambda = 0$ is a natural choice when q is small (this is in particular the configuration of [CMHR17]), not to mention that it is computionally faster. But, the proof of our theorem needs $\lambda > c_\lambda \, h_a > 0$ to hold. We think that a reasoning enabling to deal with $\lambda = 0$ should also be beneficial to the study. More generally, it would be instructive to consider a very high-dimensional setting ($p \gg n$ and not only $p \sim 10^2$ although always larger than n, as in our experiments). Such studies should follow with omic data.

Chapter 3

A BAYESIAN APPROACH FOR PARTIAL GAUSSIAN GRAPHICAL MODELS WITH SPARSITY

Gaussian graphical model is a widely studied topic both in the frequentist framework (as discussed above), and in the Bayesian one. For this second type of inference, we refer the reader for example to Chapter 10 of Maathuis *et al.* [MDLW18] where various Wishart-type priors are considered for Ω, see also Li *et al.* [LMC19] or Gan *et al.* [GYNL19] for spike-and-slab approaches and all references within. However, to the best of our knowledge, the Bayesian approach for partial Gaussian graphical models is a new research topic.

In this chapter we present the results of a collaboration with Frédéric Proïa and Pascal Jézéquel, which has been published in Bayesian Analysis in 2022 [OOJP22][1]. We explore various Bayesian approaches to estimate PGGM, and propose hierarchical structures enable to deal with single-output as well as multiple-output linear regressions, in small or high dimension, enforcing either no sparsity, sparsity, group sparsity or even sparse-group sparsity for a bi-level selection through partial correlations (direct links) between predictors and responses. Our work is inspired by the ideas of Xu and Ghosh [XG15] for the single-output setting ($q = 1$), and by the ones of Liquet *et al.* [LMPS17] for the multiple-output setting ($q > 1$). Taking advantage of the relations (1), we consider that a Gaussian prior for B must remain Gaussian for Δ (with a correctly updated variance), and that an inverse Wishart prior for R merely becomes a Wishart one for Ω_y. Yet, despite these seemingly small changes in the design of the priors, we will see that the resulting distributions are completely different. The hierarchical models that we are going to study all come from this working base, but let us point out that a wide variety of refinements exist in the recent literature for Bayesian sparsity, like the grouped 'horseshoe' of Xu *et al.* [XSM+16], the 'aggressive' multivariate Dirichlet-Laplace prior of Wei *et al.* [WRHG20], the theoretical results for group selection consistency of Yang and Narisetty [YN20] or even the extension of the Bayesian spike-and-slab group selection to generalized additive models of Bai *et al.* [BMA+20], all related to the regression setting but that might also be investigated for PGGMs. To enforce various types of sparsity in Δ for high-dimensional problems, we decided to make use of spike-and-slab priors, with a spike probability guided by a conjugate Beta distribution.

This chapter is organized as follows. Sections 3.2, 3.3 and 3.4 are dedicated to the study of our hierarchical models enforcing either no sparsity, sparsity, group sparsity or sparse-group sparsity in the

1. The codes and the dataset are available at `https://github.com/EuniceOkome/BayesPGGM`

direct links, respectively, according to the terminology of Section 2.1 of Giraud [Gir14]. In particular, we will see that our bi-level selection clearly diverges from the strategy of Liquet *et al.* [LMPS17]. We also adapt the reasoning of Yang and Narisetty [YN20] to establish group selection consistency under some technical assumptions and an appropriate amount of sparsity. Section 3.5 is devoted to the conditional posterior distributions of the parameters in order to implement Gibbs samplers that are tested in Section 3.6. This empirical section is focused on a simulation study first, to evaluate and compare the efficiency of the models, then a real dataset is treated, and a short conclusion ends the paper. But, firstly, let us give some examples of what exactly we mean by 'sparse', 'group-sparse' and 'sparse-group-sparse' settings, and let us summarize the definitions that we have chosen to retain for the well-known distributions as well as for the less usual ones, in order to avoid any misinterpretation of our results and proofs.

Example 3.0.1. To explain a set of phenotypic traits, suppose that we investigate a large collection of genetic markers spread over twenty chromosomes. For coordinate sparsity ('sparse' setting), only a few markers are active. For group sparsity ('group-sparse' setting), the markers are clustered into groups (formed by chromosomes) and only a few of them are active. For sparse-group sparsity ('sparse-group-sparse' setting), only a few chromosomes are active and they are sparse, the result is a bi-level selection (chromosomes and markers). This will be the context of our example on real data (Section 3.6.2).

3.1 Introduction

Suppose now that we observe q-dimensional matrix of responses $Y \in \mathbb{R}^{n\times q}$ where the k-th row is ${Y_i}^t$, and the p-dimensional matrix of predictors $X \in \mathbb{R}^{n\times p}$ where the k-th row is ${X_i}^t$. We are dealing with a multivariate linear regression of the form

$$Y = XB + E$$

where $B \in \mathbb{R}^{p\times q}$ contains the regression coefficients and $E \in \mathbb{R}^{n\times q}$ is a matrix-variate Gaussian noise. We are in the PGGM setting discussed in Section 1.5. In short, assuming that the couple $(Y_i, X_i) \in \mathbb{R}^{q+p}$ is jointly normally distributed with zero mean, covariance Σ and precision Ω, we can see that the blockwise decomposition given by

$$\Omega = \begin{pmatrix} \Omega_y & \Delta \\ \Delta^t & \Omega_x \end{pmatrix}$$

with $\Omega_y \in \mathbb{S}^q_{++}$, $\Delta \in \mathbb{R}^{q\times p}$ and $\Omega_x \in \mathbb{S}^p_{++}$ leads to $Y_i \,|\, X_i \sim \mathcal{N}_q(-\Omega_y^{-1}\Delta\, X_i, \Omega_y^{-1})$. This is a crucial remark because one can see that the multiple-output regression $Y_i = B^t X_i + E_i$ with Gaussian noise $E_i \sim \mathcal{N}_q(0, R)$ may be reparametrized with

$$B = -\Delta^t \Omega_y^{-1} \quad \text{and} \quad R = \Omega_y^{-1}. \tag{1}$$

Before going deeper into the subject, let us summarize the definitions that we have chosen to retain for the well-known distributions as well as for the less usual ones, in order to avoid any misinterpretation of our results and proofs.

Definition 3.1.1 (Gaussian)**.** The density of $X \in \mathbb{R}^{d_1 \times d_2}$ following the matrix normal distribution $\mathcal{MN}_{d_1 \times d_2}(M, \Sigma_1, \Sigma_2)$ is given by

$$p(X) = \frac{1}{(2\pi)^{\frac{d_1 d_2}{2}} |\Sigma_1|^{\frac{d_2}{2}} |\Sigma_2|^{\frac{d_1}{2}}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(\Sigma_2^{-1}(X-M)^t \Sigma_1^{-1}(X-M)\right)\right)$$

where $M \in \mathbb{R}^{d_1 \times d_2}$, $\Sigma_1 \in \mathbb{S}_{++}^{d_1}$ and $\Sigma_2 \in \mathbb{S}_{++}^{d_2}$. When $d_2 = 1$, this is a multivariate normal distribution $\mathcal{N}_d(\mu, \Sigma)$ with $d = d_1$, $\mu = M$ and $\Sigma = \Sigma_2^{-1}\Sigma_1$, having density

$$p(X) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (X-\mu)^t \Sigma^{-1}(X-\mu)\right)$$

where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{S}_{++}^d$.

Definition 3.1.2 (Generalized Inverse Gaussian)**.** The density of $X \in \mathbb{S}_{++}^d$ following the matrix generalized inverse Gaussian distribution $\mathcal{MGIG}_d(\nu, A, B)$ is given by

$$p(X) = \frac{|X|^{\nu - \frac{d+1}{2}}}{\left|\frac{A}{2}\right|^{\nu} B_\nu\left(\frac{A}{2}, \frac{B}{2}\right)} \exp\left(-\frac{1}{2} \operatorname{tr}\left(A X^{-1} + B X\right)\right) \mathbb{1}_{\{X \in \mathbb{S}_{++}^d\}}$$

where $\nu \in \mathbb{R}$, $A \in \mathbb{S}_{++}^d$, $B \in \mathbb{S}_{++}^d$ and B_ν is a Bessel-type function of order ν. When $d = 1$, this is a generalized inverse Gaussian distribution $\mathcal{GIG}(\nu, a, b)$ with $a = A$ and $b = B$, having density

$$p(X) = \frac{X^{\nu-1}}{\left(\frac{a}{2}\right)^{\nu} B_\nu\left(\frac{a}{2}, \frac{b}{2}\right)} \mathrm{e}^{-\frac{a}{2X} - \frac{bX}{2}} \mathbb{1}_{\{X > 0\}}$$

where $\nu \in \mathbb{R}$, $a > 0$ and $b > 0$.

Definition 3.1.3 (Wishart/Gamma/Exponential)**.** The density of $X \in \mathbb{S}_{++}^d$ following the matrix Wishart distribution $\mathcal{W}_d(u, V)$ is given by

$$p(X) = \frac{|X|^{\frac{u-d-1}{2}}}{2^{\frac{du}{2}} \Gamma_d\left(\frac{u}{2}\right) |V|^{\frac{u}{2}}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(V^{-1}X\right)\right) \mathbb{1}_{\{X \in \mathbb{S}_{++}^d\}}$$

where $u > d - 1$, $V \in \mathbb{S}_{++}^d$ and Γ_d is the multivariate Gamma function of order d. When $d = 1$, this is a Gamma distribution $\Gamma(a, b)$ with $a = \frac{u}{2}$ and $\frac{1}{b} = 2V$, having density

$$p(X) = \frac{b^a X^{a-1}}{\Gamma(a)} \mathrm{e}^{-bX} \mathbb{1}_{\{X > 0\}}$$

where $a > 0$ and $b > 0$. The exponential distribution $\mathcal{E}(\ell)$ is then defined as the $\Gamma(1, \ell)$ distribution, for $\ell > 0$.

Definition 3.1.4 (Beta)**.** The density of $X \in [0, 1]$ following the Beta distribution $\beta(a, b)$ is given by

$$p(X) = \frac{X^{a-1}(1-X)^{b-1}}{\beta(a, b)} \mathbb{1}_{\{0 \leqslant X \leqslant 1\}}$$

where $a > 0$, $b > 0$ and β is the Beta function.

In all the paper, data and parameters are gathered in $\Theta = \{Y, X, \Delta, \Omega_y, \nu, \lambda, \pi\}$ and, to standardize, for any $e \in \Theta$, we note $\Theta_e = \Theta \backslash \{e\}$.

3.2 The sparse setting

In this section, $\lambda_i \in \mathbb{R}$ is the i-th component of $\lambda \in \mathbb{R}^p$, $\Delta_i \in \mathbb{R}^q$ is the i-th column of Δ and $X_i \in \mathbb{R}^n$ stand for the i-th column of X ($1 \leqslant i \leqslant p$). Let us consider the hierarchical Bayesian model, where the columns of Δ are assumed to be independent, given by

$$\left\{ \begin{array}{lcl} Y \,|\, X, \Delta, \Omega_y & \sim & \mathcal{MN}_{n\times q}(-X\,\Delta^t\,\Omega_y^{-1},\, I_n,\, \Omega_y^{-1}) \\ \Delta_i \,|\, \Omega_y, \lambda_i, \pi & \overset{\perp\!\!\!\perp}{\sim} & (1-\pi)\,\mathcal{N}_q(0,\, \lambda_i\,\Omega_y) + \pi\,\delta_0 \\ \lambda_i & \overset{\perp\!\!\!\perp}{\sim} & \Gamma(\alpha,\, \ell_i) \\ \Omega_y & \sim & \mathcal{W}_q(u,\, V) \\ \pi & \sim & \beta(a,\, b) \end{array} \right. \tag{2}$$

for $i \in [\![1, p]\!]$, with hyperparameters $\alpha = \frac{1}{2}(q+1)$, $\ell_i > 0$, $u > q-1$, $V \in \mathbb{S}_{++}^q$, $a > 0$ and $b > 0$. A general ungrouped sparsity is promoted in the columns of Δ through the spike-and-slab prior. In this mixture model, π is the prior spike probability and λ is an adaptative shrinkage factor acting at the predictor scale (λ_i is associated with the direct links between predictor i and all the responses). When $\ell_i = \ell$ for all i, we will rather speak of global shrinkage. The degree of sparsity will be characterized by the number N_0 of zero columns of Δ, that is

$$N_0 = \mathrm{Card}(i,\, \Delta_i = 0) = \sum_{i=1}^{p} \mathbb{1}_{\{\Delta_i = 0\}}. \tag{3}$$

To implement a Gibbs sampler from the full posterior distribution stemming from (2), we may use the conditional distributions given in the proposition below.

Proposition 3.2.1. *In the hierarchical model* (2)*, the conditional posterior distributions are as follows.*

– *The parameter* Δ *satisfies, for* $i \in [\![1, p]\!]$*,*

$$\Delta_i \,|\, \Theta_{\Delta_i} \;\sim\; (1-p_i)\,\mathcal{N}_q(-s_i\,H_i,\, s_i\,\Omega_y) + p_i\,\delta_0$$

where

$$H_i = \Omega_y\, Y^{\,t}\, X_i + \sum_{j \neq i} \langle X_i,\, X_j \rangle\, \Delta_j, \qquad s_i = \frac{\lambda_i}{1 + \lambda_i\, \|X_i\|^{\,2}}$$

and

$$p_i = \frac{\pi}{\pi + (1-\pi)\,(1 + \lambda_i\,\|X_i\|^{\,2})^{-\frac{q}{2}}\, \exp\Big(\frac{s_i\, H_i^{\,t}\, \Omega_y^{-1} H_i}{2}\Big)}.$$

– *The parameter* Ω_y *satisfies*

$$\Omega_y \,|\, \Theta_{\Omega_y} \;\sim\; \mathcal{MGIG}_q\Big(\frac{n-p+N_0+u}{2},\, \Delta\,(X^t\,X + D_\lambda^{-1})\,\Delta^t,\, Y^t\,Y + V^{-1}\Big)$$

where $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$.

– *The parameter* λ *satisfies, for* $i \in [\![1, p]\!]$,

$$\lambda_i \,|\, \Theta_{\lambda_i} \;\sim\; \mathbb{1}_{\{\Delta_i \neq 0\}}\, \mathcal{GIG}\Big(\frac{1}{2},\, \Delta_i^t\,\Omega_y^{-1}\Delta_i,\, 2\,\ell_i\Big) + \mathbb{1}_{\{\Delta_i = 0\}}\,\Gamma(\alpha,\, \ell_i).$$

– *The parameter* π *satisfies*

$$\pi \,|\, \Theta_\pi \;\sim\; \beta\big(N_0 + a,\, p - N_0 + b\big).$$

Proof. See Section 3.5.1. □

Remark 3.2.2. The Bayesian Lasso, as introduced *e.g.* in Section 6.1 of [HTW15] or in [PC08], assumes a prior Laplace distribution for the regression coefficients conditional on the noise variance. In our case, $\Delta_i \,|\, \Omega_y, \pi$ is still a multivariate spike-and-slab (after integrating over λ_i), with a slab following a so-called multivariate K-distribution (see [EKL06]), which is a generalization of the multivariate Laplace distribution. See *e.g.* Section 2.1 of [LMPS17]. From this point of view, our study is in line with the usual Bayesian regression schemes. Perhaps even more interesting, going on with the idea of the authors, suppose that, for all $1 \leqslant i \leqslant p$, $\Delta_i = b_i\,\Delta_i^*$ where Δ_i^* follows the multivariate K-distribution described above and $b_i \,|\, \pi \sim \mathcal{B}(1-\pi)$ is independent of Δ_i^*. Now, the sparsity in Δ is not induced by a spike-and-slab strategy anymore but, equivalently, by multiplying the slab part by an independent Bernoulli variable being 0 with probability π. Then, it is possible to show that the negative log-likelihood of this alternative hierarchical model is given, up to an additive constant that does not depend on Δ, by

$$\frac{1}{2}\,\Big\|(Y + X\,\Delta^t\,\Omega_y^{-1})\,\Omega_y^{\frac{1}{2}}\Big\|_F^2 + \sum_{i=1}^p c_i\,\Big\|\Omega_y^{-\frac{1}{2}}\Delta_i^*\Big\|_F + \ln\Big(\frac{1-\pi}{\pi}\Big)\sum_{i=1}^p b_i$$

where $c_i > 0$. We first recognize an ℓ_2-type penalty but also an ℓ_0-type penalty on Δ (provided that $\pi < \frac{1}{2}$) since summing the b_i amounts to counting the number of non-zero columns in Δ. Consequently, there is a close connection between our hierarchical Bayesian model and the regressions penalized by ℓ_2 and ℓ_0 norms, problems that are known to be very hard to solve due to combinatorial optimization.

The particular case $q = 1$ is a very useful corollary of the proposition. Here, the direct links form a row vector such that $\Delta^t \in \mathbb{R}^p$ with components $\Delta_i \in \mathbb{R}$ $(1 \leqslant i \leqslant p)$, and the precision matrix of the responses reduces to $\omega_y > 0$. According to the parametrization of the distributions (see Section 3.1), the corresponding prior distribution of ω_y is $\Gamma(\frac{u}{2}, \frac{1}{2\,v})$ for $u, v > 0$ and the one of λ_i is $\mathcal{E}(\ell_i)$ for $\ell_i > 0$. The other priors are unchanged.

Corollary 3.2.3. *In the hierarchical model* (2) *with* $q = 1$*, the conditional posterior distributions are as follows.*

– *The parameter* Δ *satisfies, for* $i \in [\![1, p]\!]$,

$$\Delta_i \,|\, \Theta_{\Delta_i} \;\sim\; (1-p_i)\,\mathcal{N}(-s_i\,h_i,\, s_i\,\omega_y) + p_i\,\delta_0$$

where

$$h_i = \omega_y\,\langle X_i,\, Y\rangle + \sum_{j\,\neq\, i}\langle X_i,\, X_j\rangle\,\Delta_j, \qquad s_i = \frac{\lambda_i}{1+\lambda_i\,\|X_i\|^{\,2}}$$

and

$$p_i = \frac{\pi}{\pi + (1-\pi)\,(1+\lambda_i\,\|X_i\|^{\,2})^{-\frac{1}{2}}\,\exp\!\Big(\frac{s_i\,h_i^2}{2\,\omega_y}\Big)}.$$

– *The parameter* ω_y *satisfies*

$$\omega_y \,|\, \Theta_{\omega_y} \;\sim\; \mathcal{GIG}\!\left(\frac{n-p+N_0+u}{2},\, \Delta\,(X^t\,X + D_\lambda^{-1})\,\Delta^t,\, \|Y\|^{\,2} + \frac{1}{v}\right)$$

where $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$.

– *The parameter* λ *satisfies, for* $i \in [\![1, p]\!]$,

$$\lambda_i \,|\, \Theta_{\lambda_i} \;\sim\; \mathbb{1}_{\{\Delta_i\,\neq\,0\}}\,\mathcal{GIG}\!\left(\frac{1}{2},\, \frac{\Delta_i^2}{\omega_y},\, 2\,\ell_i\right) + \mathbb{1}_{\{\Delta_i\,=\,0\}}\,\mathcal{E}(\ell_i).$$

– *The parameter* π *satisfies*

$$\pi \,|\, \Theta_\pi \;\sim\; \beta\big(N_0 + a,\, p - N_0 + b\big).$$

Proof. This is a consequence of Proposition 3.2.1. □

Note that we can also easily derive the Bayesian counterpart of the standard PGGM adapted to the small-dimensional case, with no sparsity, by taking $\pi = 0$.

Corollary 3.2.4. *In the hierarchical model* (2) *with* $\pi = 0$*, the conditional posterior distributions are as follows.*

– *The parameter* Δ *satisfies, for* $i \in [\![1, p]\!]$,

$$\Delta_i \,|\, \Theta_{\Delta_i} \;\sim\; \mathcal{N}_q(-s_i\,H_i,\, s_i\,\Omega_y)$$

where

$$H_i = \Omega_y\,Y^{\,t}\,X_i + \sum_{j\,\neq\, i}\langle X_i,\, X_j\rangle\,\Delta_j \qquad \textit{and} \qquad s_i = \frac{\lambda_i}{1+\lambda_i\,\|X_i\|^{\,2}}.$$

– *The parameter* Ω_y *satisfies*

$$\Omega_y \,|\, \Theta_{\Omega_y} \;\sim\; \mathcal{MGIG}_q\!\left(\frac{n-p+u}{2},\, \Delta\,(X^t\,X + D_\lambda^{-1})\,\Delta^t,\, Y^{\,t}\,Y + V^{-1}\right)$$

where $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$.

– *The parameter λ satisfies, for $i \in [\![1, p]\!]$,*

$$\lambda_i \,|\, \Theta_{\lambda_i} \;\sim\; \mathcal{GIG}\Big(\frac{1}{2},\, \Delta_i^t\, \Omega_y^{-1} \Delta_i,\, 2\,\ell_i\Big).$$

Proof. This is a consequence of Proposition 3.2.1. □

In the simulation study of Section 3.6.1, Scen. 0, 1 and 2 are dedicated to the sparse setting. The next section discusses the group sparsity in Δ.

3.3 The group-sparse setting

The predictors are now ordered in m groups of sizes $\kappa_1 + \ldots + \kappa_m = p$. For the g-th group ($1 \leqslant g \leqslant m$), $\lambda_g \in \mathbb{R}$ is the g-th component of $\lambda \in \mathbb{R}^m$, the covariate submatrix is $\underset{\sim}{X}_g \in \mathbb{R}^{n \times \kappa_g}$ and the corresponding slice of Δ is $\underset{\sim}{\Delta}_g \in \mathbb{R}^{q \times \kappa_g}$. Let us consider the hierarchical Bayesian model, where the columns of Δ are assumed to be independent both within and between the groups, given by

$$\left\{\begin{array}{lcl} Y \,|\, X, \Delta, \Omega_y & \sim & \mathcal{MN}_{n\times q}(-X\,\Delta^t\,\Omega_y^{-1},\, I_n,\, \Omega_y^{-1}) \\ \underset{\sim}{\Delta}_g \,|\, \Omega_y, \lambda_g, \pi & \overset{\perp\!\!\!\perp}{\sim} & (1-\pi)\,\mathcal{MN}_{q\times\kappa_g}(0,\, \lambda_g\,\Omega_y, I_{\kappa_g}) + \pi\,\delta_0 \\ \lambda_g & \overset{\perp\!\!\!\perp}{\sim} & \Gamma(\alpha_g,\, \ell_g) \\ \Omega_y & \sim & \mathcal{W}_q(u,\, V) \\ \pi & \sim & \beta(a,\, b) \end{array}\right. \tag{4}$$

for $g \in [\![1, m]\!]$, with hyperparameters $\alpha_g = \frac{1}{2}(q\,\kappa_g + 1)$, $\ell_g > 0$, $u > q - 1$, $V \in \mathbb{S}_{++}^q$, $a > 0$ and $b > 0$. A general group sparsity is promoted in the columns of Δ through the spike-and-slab prior at the group level. In this mixture model, π is the prior spike probability and λ is an adaptative shrinkage factor acting at the group scale (λ_g is associated with the direct links between the predictors of group g and all the responses). Likewise, when $\ell_g = \ell$ for all g, we will rather speak of global shrinkage. Now, the degree of sparsity will be characterized by N_0 given in (3), but also by the number G_0 of zero groups of Δ, that is

$$G_0 = \mathrm{Card}(g,\, \underset{\sim}{\Delta}_g = 0) = \sum_{g=1}^{m} \mathbb{1}_{\{\underset{\sim}{\Delta}_g = 0\}}. \tag{5}$$

To implement a Gibbs sampler from the full posterior distribution stemming from (4), we may use the conditional distributions given in the proposition below.

Proposition 3.3.1. *In the hierarchical model* (4)*, the conditional posterior distributions are as follows.*

– *The parameter Δ satisfies, for $g \in [\![1, m]\!]$,*

$$\underset{\sim}{\Delta}_g \,|\, \Theta_{\Delta_g} \;\sim\; (1-p_g)\,\mathcal{MN}_{q\times\kappa_g}(-H_g\,S_g,\, \Omega_y,\, S_g) + p_g\,\delta_0$$

where

$$H_g = \Omega_y Y^{\,t}\,\underset{\sim}{X}_g + \sum_{j\,\neq\, g} \underset{\sim}{\Delta}_j\,\underset{\sim}{X}_j^{\,t}\,\underset{\sim}{X}_g, \qquad S_g = \lambda_g\,\big(I_{\kappa_g} + \lambda_g\,\underset{\sim}{X}_g^{\,t}\,\underset{\sim}{X}_g\big)^{-1}$$

and

$$p_g = \frac{\pi}{\pi + (1-\pi)\,|I_{\kappa_g} + \lambda_g\, \underline{X}_g^t\, \underline{X}_g|^{-\frac{q}{2}}\, \exp\Big(\frac{\mathrm{tr}(H_g^t\, \Omega_y^{-1} H_g\, S_g)}{2}\Big)}.$$

– *The parameter Ω_y satisfies*

$$\Omega_y \,|\, \Theta_{\Omega_y} \;\sim\; \mathcal{MGIG}_q\Big(\frac{n-p+N_0+u}{2},\, \Delta\,(X^t\,X + D_\lambda^{-1})\,\Delta^t,\, Y^t\,Y + V^{-1}\Big)$$

where $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_1, \ldots, \lambda_m, \ldots, \lambda_m)$ with each λ_g duplicated κ_g times.

– *The parameter λ satisfies, for $g \in [\![1, m]\!]$,*

$$\lambda_g \,|\, \Theta_{\lambda_g} \;\sim\; \mathbb{1}_{\{\Delta_g \neq 0\}}\, \mathcal{GIG}\Big(\frac{1}{2},\, \mathrm{tr}(\underline{\Delta}_g^t\, \Omega_y^{-1} \underline{\Delta}_g),\, 2\,\ell_g\Big) + \mathbb{1}_{\{\Delta_g = 0\}}\, \Gamma(\alpha_g,\, \ell_g).$$

– *The parameter π satisfies*

$$\pi \,|\, \Theta_\pi \;\sim\; \beta\big(G_0 + a,\, m - G_0 + b\big).$$

Proof. See Section 3.5.2. □

Note that Remark 3.2.2 still applies to this configuration, after some adjustments (the ℓ_0-like penalty is on the number of non-zero groups). Here again, the particular case $q = 1$ is a very useful corollary. The direct links form a row vector such that $\Delta^t \in \mathbb{R}^p$ with groups $\underline{\Delta}_g^t \in \mathbb{R}^{\kappa_g}$ $(1 \leqslant g \leqslant m)$, the precision matrix of the responses reduces to $\omega_y > 0$. According to the parametrization of the distributions (see Section 3.1), the corresponding prior distribution of ω_y is $\Gamma(\frac{u}{2}, \frac{1}{2\,v})$ for $u, v > 0$, like in the ungrouped setting. The other priors are unchanged.

Corollary 3.3.2. *In the hierarchical model* (4) *with $q = 1$, the conditional posterior distributions are as follows.*

– *The parameter Δ satisfies, for $g \in [\![1, m]\!]$,*

$$\underline{\Delta}_g^t \,|\, \Theta_{\Delta_g} \;\sim\; (1-p_g)\,\mathcal{N}_{\kappa_g}(-S_g\, H_g,\, \omega_y\, S_g) + p_g\, \delta_0$$

where

$$H_g = \omega_y\, \underline{X}_g^t\, Y + \sum_{j \neq g} \underline{X}_g^t\, \underline{X}_j\, \underline{\Delta}_j^t, \qquad S_g = \lambda_g\, \big(I_{\kappa_g} + \lambda_g\, \underline{X}_g^t\, \underline{X}_g\big)^{-1}$$

and

$$p_g = \frac{\pi}{\pi + (1-\pi)\,|I_{\kappa_g} + \lambda_g\, \underline{X}_g^t\, \underline{X}_g|^{-\frac{1}{2}}\, \exp\Big(\frac{H_g^t\, S_g\, H_g}{2\,\omega_y}\Big)}.$$

– *The parameter ω_y satisfies*

$$\omega_y \,|\, \Theta_{\omega_y} \;\sim\; \mathcal{GIG}\Big(\frac{n-p+N_0+u}{2},\, \Delta\,(X^t\,X + D_\lambda^{-1})\,\Delta^t,\, \|Y\|^2 + \frac{1}{v}\Big)$$

where $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_1, \ldots, \lambda_m, \ldots, \lambda_m)$ with each λ_g duplicated κ_g times.

– *The parameter λ satisfies, for $g \in [\![1, m]\!]$,*

$$\lambda_g \,|\, \Theta_{\lambda_g} \;\sim\; \mathbb{1}_{\{\Delta_g \neq 0\}}\, \mathcal{GIG}\Big(\frac{1}{2},\, \frac{\|\underline{\Delta}_g\|^2}{\omega_y},\, 2\,\ell_g\Big) + \mathbb{1}_{\{\Delta_g = 0\}}\, \Gamma(\alpha_g,\, \ell_g).$$

– *The parameter π satisfies*

$$\pi \,|\, \Theta_\pi \;\sim\; \beta\big(G_0 + a,\, m - G_0 + b\big).$$

Proof. This is a consequence of Proposition 3.3.1. □

In the simulation study of Section 3.6.1, Scen. 3 and 4 are dedicated to the group-sparse setting. To conclude this section, a theoretical guarantee is provided (given Ω_y and with $\lambda = \lambda_n$ and $\pi = \pi_n$ depending on n). It is possible to obtain a model selection consistency property for this approach when both the number of observations n and the number of groups $m = m_n$ tend to infinity, by adapting the reasoning of [YN20] dedicated to the linear regression (with $q = 1$). Indeed, when Ω_y is known, Δ reduces to a linear transformation of B. Thus, it is not surprising that a similar result follows under the same kind of hypotheses. In the sequel, we denote by $X_{(k)} \in \mathbb{R}^{n \times |k|}$ the design matrix of rank r_k corresponding to the submodel indexed by the binary vector $k \in \{0,1\}^m$ having $|k|$ non-zero values ($k_g = 1$ means that the g-th group is included in the model), and by $\Pi_{(k)} \in \mathbb{R}^{n \times n}$ the projection matrix onto the column-space of $X_{(k)}$. Similarly, Δ restricted to k is $\Delta_{(k)} \in \mathbb{R}^{q \times |k|}$. The true model is called t and $t^{\pm g}$ are submodels of t that contain only the g-th group or that are deprived of it, respectively. Let

$$\delta_1 = \inf_{1 \leqslant g \leqslant |t|} \big\| (I_n - \Pi_{(t^{-g})})\, X_{(t^{+g})}\, \Delta^t_{(t^{+g})}\, \Omega_y^{-\frac{1}{2}} \big\|_F^2$$

and, for some $K > 0$,

$$\delta_2^K = \inf_{k \,\in\, E_i} \big\| (I_n - \Pi_{(k)})\, X_{(t)}\, \Delta^t_{(t)}\, \Omega_y^{-\frac{1}{2}} \big\|_F^2$$

with $E_i = \{k \text{ such that } t \not\subset k \text{ and } r_k \leqslant K r_t\}$. Let also,

$$\mu^K_{n,\,\min} = \inf_{k \,\in\, F_K} \mu^+\!\left(\frac{X^t_{(k)}\, X_{(k)}}{n}\right) \quad \text{and} \quad \bar{\mu}_n = \inf_{k \,\in\, F} \mu^*\!\left(\frac{X^t_{(k)}\, (I_n - \Pi_{(k \cap t)})\, X_{(k)}}{n}\right)$$

with $F_K = \{k \text{ such that } t \subset k \text{ and } r_k \leqslant (K+1)\, r_t\}$ and $F = \{k \text{ such that } |k \backslash t| > 0\}$, and where, for a square matrix A, $\mu^+(A)$ is the minimum non-zero eigenvalue of A and $\mu^*(A)$ is the geometric mean of the non-zero eigenvalues of A. The hypotheses are those of [YN20] that we have to slightly adapt. By $f_n \asymp g_n$ we mean that there is a constant $c \neq 0$ such that $f_n / g_n \to c$ as n tends to infinity.

(H.1) There exists a rate such that $m_n = \mathrm{e}^{v_n}$ with $v_n \to +\infty$ and $v_n = o(n)$.

(H.2) The prior slab probability satisfies $1 - \pi_n \asymp 1/m_n$.

(H.3) The shrinkage factors satisfy $n\lambda_n^\sharp \asymp m_n^{2+\eta}\, \bar{\mu}_n^{-\eta}$ and $\mu^K_{n,\,\min}\, n\lambda_n^\sharp \to +\infty$ for some $\eta > 0$, where $\lambda_n^\sharp = \max_i \lambda_{n,\,i}$.

(H.4) There exists $\epsilon_1 > 0$ such that $\delta_1 > (1+\epsilon_1)\, r_t\, [(4+\eta) \ln m_n - \eta \ln \bar{\mu}_n]$.

(H.5) There exists $\epsilon_2 > 0$ such that $\delta_2^K > (1+\epsilon_2)\, r_t \left[(4+\eta)\ln m_n - \eta \ln \bar{\mu}_n\right]$ for some $K > \max(8/\eta + 1, \eta/(\eta-1))$.

We refer the reader to p. 917 of [YN20] where the authors give very clarifying comments on the interpretation to be given to these technical assumptions. In particular, while (H.1), (H.2) and (H.3) control the behavior of m_n, π_n and λ_n as n tends to infinity, (H.4) and (H.5) are related to sensitivity and specificity and are therefore in connection with the true model t.

Proposition 3.3.3. *Suppose that (H.1)–(H.5) are satisfied. Then, as n tends to infinity,*

$$\mathbb{P}(\mathcal{T} \,|\, Y, X, \Omega_y) \xrightarrow{\mathbb{P}} 1$$

where $\mathcal{T} = \{t \text{ is selected}\}$ and t is the true model.

Proof. The result is obtained by following the same lines as the proof of Theorem 2.1 of [YN20]. One just has to clarify a few points to solve the issues arising from $q \geqslant 1$ and from the adaptative shrinkage, which is done in Section 3.5.4. □

Remark 3.3.4. Obviously, Proposition 3.3.3 also holds for the sparse setting (with $m = p$) and in that case, it is instructive to draw the parallel with Theorem 1 of [RSZZ15] even if the estimation procedure is very different. The authors show that, to obtain a $\sqrt{n}$-consistent estimation of the precision matrix Ω in a GGM, Ω must contain at most $\asymp \sqrt{n}/\ln p$ non-zero columns. In the Gibbs sampler (see Proposition 3.2.1), the slab probability $1-\pi$ is generated according to a distribution that satisfies

$$\mathbb{E}[1-\pi \,|\, \Theta_\pi] = \frac{p - N_0 + b}{p + a + b} \quad \text{and} \quad \mathbb{V}(1-\pi \,|\, \Theta_\pi) = \frac{(N_0 + a)(p - N_0 + b)}{(p+a+b)^2\,(p+a+b+1)}.$$

Thus, if the model selects $\asymp \sqrt{n}/\ln p$ predictors, it follows that the posterior expectation of $1-\pi$ is $\asymp \sqrt{n}/(p \ln p) = 1/p$ when $p = \mathrm{e}^{\sqrt{n}}$. In that case, the posterior variance of $1-\pi$ is $\asymp 1/p^2$. To sum up, in a model with $\asymp \sqrt{n}/\ln p$ predictors selected, the posterior distribution of $1-\pi$ is very concentrated around $1/p$ which conforms to (H.1) and (H.2). This is not directly comparable due to the different procedures, but it seems interesting to observe that the same orders of magnitude are involved to reach theoretical guarantees for the estimation of Δ.

In the next section, an approach is suggested to deal with sparse-group sparsity in Δ, for a bi-level selection.

3.4 The sparse-group-sparse setting

To produce a sparse model both at the variable level (for variable selection) and at the group level (for group selection), it seems natural to carry on with our strategy by introducing another spike-and-slab effect into the first one. The predictors are still ordered in m groups of sizes $\kappa_1 + \ldots + \kappa_m = p$. For the g-th group ($1 \leqslant g \leqslant m$), $\lambda_g \in \mathbb{R}$ is the g-th component of $\lambda \in \mathbb{R}^m$ and, for the i-th predictor of this group ($1 \leqslant i \leqslant \kappa_g$), $\nu_{gi} \in \mathbb{R}$ is the i-th component of $\nu_g \in \mathbb{R}^{\kappa_g}$. The i-th column of the covariate submatrix $\underset{\sim}{X}_g$ is $X_{gi} \in \mathbb{R}^n$ and the corresponding slice of $\underset{\sim}{\Delta}_g$ is $\Delta_{gi} \in \mathbb{R}^q$ while $\underset{\sim}{\Delta}_{g\setminus i} \in \mathbb{R}^{q \times (\kappa_g - 1)}$ is $\underset{\sim}{\Delta}_g$ deprived

of Δ_{gi}. Here our approach diverges from [XG15] and [LMPS17]. The bi-level selection of the authors is made through spike-and-slab effects both at the group scale and on the individual variances, considered as truncated Gaussians, generating zero groups and (almost surely) zero coefficients within the groups. Let us suggest instead the Bayesian hierarchical model given by

$$\left\{\begin{array}{lcl} Y \,|\, X, \Delta, \Omega_y & \sim & \mathcal{MN}_{n\times q}(-X\,\Delta^t\,\Omega_y^{-1},\, I_n,\, \Omega_y^{-1}) \\ \underline{\Delta}_g \,|\, \nu_g, \lambda_g, \pi & \overset{\perp\!\!\!\perp}{\sim} & (1-\pi_1)\left[(1-\pi_2)\,\mathcal{N}_q(0,\, \lambda_g\,\nu_{gi}\,\Omega_y) + \pi_2\,\delta_0\right]^{\otimes \kappa_g} + \pi_1\,\delta_0 \\ \nu_{gi} & \overset{\perp\!\!\!\perp}{\sim} & \Gamma(\alpha,\, \ell_{gi}) \\ \lambda_g & \overset{\perp\!\!\!\perp}{\sim} & \Gamma(\alpha_g,\, \gamma_g) \\ \Omega_y & \sim & \mathcal{W}_q(u,\, V) \\ \pi_j & \overset{\perp\!\!\!\perp}{\sim} & \beta(a_j,\, b_j) \end{array}\right. \tag{6}$$

for $g \in [\![1, m]\!]$, $i \in [\![1, \kappa_g]\!]$ and $j \in [\![1, 2]\!]$, with hyperparameters $\alpha = \frac{1}{2}(q+1)$, $\alpha_g = \frac{1}{2}(q\,\kappa_g + 1)$, $\ell_{gi} > 0$, $\gamma_g > 0$, $u > q-1$, $V \in \mathbb{S}_{++}^q$, $a_j > 0$, and $b_j > 0$. In this mixture model, π_1 is the prior spike probability on the groups whereas π_2 is the prior spike probability within the non-zero groups, for a bi-level selection. In terms of cumulative shrinkage effects, λ is an adaptative shrinkage factor acting at the group scale and ν is an adaptative shrinkage factor acting at the predictor scale (λ_g is associated with the direct links between the predictors of group g and all the responses whereas ν_{gi} is associated with the direct links between predictor i of group g and all the responses). In this way, (6) opens up many perspectives for dealing with bi-level shrinkage. We can set $\gamma_g = \gamma$ for all g, for a global shrinkage at the group scale. At the predictor scale, when $\ell_{gi} = \ell_g$ for all i, this is a global shrinkage in the g-th group but we might even consider a full global shrinkage $\ell_{gi} = \ell$. However, an identifiability issue may result from the product $\lambda_g\,\nu_{gi}$ between group and within-group effects. Even if the posterior distributions depend on different levels of data that shall resolve it, one can for example fix $\lambda_g = 1$ (for adaptative) or $\nu_{gi} = 1$ (for global) and let the shrinkage entirely rely on the other parameter. Although it achieves the same objectives as those of [XG15] and [LMPS17], this hierarchy seems more consistent with our previous sections (take $\pi_2 = 0$ and $\nu_{gi} = 1$ to remove the within-group effect and recover the group-sparse setting of Section 3.3, take $\pi_1 = 0$ and $\lambda_g = 1$ to remove the group effect and recover the sparse setting of Section 3.2). In this context, the degree of sparsity is still characterized by N_0 given in (3) for the predictor scale, by G_0 given in (5) for the group scale, but also, for the within-group scale, by the number N_{0g} of zero columns in each particular group g, that is, for all $1 \leqslant g \leqslant m$,

$$N_{0g} = \mathrm{Card}(i,\, \Delta_{gi} = 0) = \sum_{i=1}^{\kappa_g} \mathbb{1}_{\{\Delta_{gi}\,=\,0\}}. \tag{7}$$

We also need to define the number J_0 of zero columns in the non-zero groups, that is

$$J_0 = \mathrm{Card}(i,\, \Delta_{gi} = 0 \text{ and } \underline{\Delta}_g \neq 0) = \sum_{g=1}^{m} N_{0g}\,\mathbb{1}_{\{\underline{\Delta}_g\,\neq\,0\}}. \tag{8}$$

To implement a Gibbs sampler from the full posterior distribution stemming from (6), we may use the conditional distributions given in the proposition below.

Proposition 3.4.1. *In the hierarchical model* (6)*, the conditional posterior distributions are as follows.*

– *The parameter* Δ_{gi} *satisfies, for* $g \in [\![1, m]\!]$ *and* $i \in [\![1, \kappa_g]\!]$,

$$\Delta_{gi} \,|\, \Theta_{\Delta_{gi}} \;\sim\; (1 - p_{gi})\, \mathcal{N}_q(-s_{gi}\, H_{gi},\, s_{gi}\, \Omega_y) + p_{gi}\, \delta_0$$

where

$$H_{gi} = \Omega_y\, Y^{\,t}\, X_{gi} + \sum_{h,j \neq g,i} \langle X_{gi},\, X_{hj} \rangle\, \Delta_{hj}, \qquad s_{gi} = \frac{\nu_{gi}\, \lambda_g}{1 + \nu_{gi}\, \lambda_g\, \|X_{gi}\|^2}$$

and

$$p_{gi} = \frac{\rho_{gi}}{\rho_{gi} + (1 - \pi_1)\,(1 - \pi_2)\,(1 + \nu_{gi}\, \lambda_g\, \|X_{gi}\|^2)^{-\frac{q}{2}}\, \exp\!\Big(\frac{s_{gi}\, H_{gi}^t\, \Omega_y^{-1}\, H_{gi}}{2}\Big)}$$

in which $\rho_{gi} = (1 - \pi_1)\, \pi_2\, \mathbb{1}_{\{\Delta_{g\setminus i} \neq 0\}} + \pi_1\, \mathbb{1}_{\{\Delta_{g\setminus i} = 0\}}$.

– *The parameter* Ω_y *satisfies*

$$\Omega_y \,|\, \Theta_{\Omega_y} \;\sim\; \mathcal{MGIG}_q\Big(\frac{n - p + N_0 + u}{2},\, \Delta\,(X^t\, X + D_{\lambda\nu}^{-1})\,\Delta^t,\, Y^{\,t}\, Y + V^{-1}\Big)$$

where $D_{\lambda\nu} = \mathrm{diag}(\nu_{11}\, \lambda_1, \ldots, \nu_{1\kappa_1}\, \lambda_1, \ldots, \nu_{m1}\, \lambda_m, \ldots, \nu_{m\kappa_m}\, \lambda_m)$.

– *The parameter* ν *satisfies, for* $g \in [\![1, m]\!]$ *and* $i \in [\![1, \kappa_g]\!]$,

$$\nu_{gi} \,|\, \Theta_{\nu_{gi}} \;\sim\; \mathbb{1}_{\{\Delta_{gi} \neq 0\}}\, \mathcal{GIG}\left(\frac{1}{2},\, \frac{\Delta_{gi}^t\, \Omega_y^{-1} \Delta_{gi}}{\lambda_g},\, 2\,\ell_{gi}\right) + \mathbb{1}_{\{\Delta_{gi} = 0\}}\, \Gamma(\alpha,\, \ell_{gi}).$$

– *The parameter* λ *satisfies, for* $g \in [\![1, m]\!]$,

$$\lambda_g \,|\, \Theta_{\lambda_g} \;\sim\; \mathbb{1}_{\{\Delta_g \neq 0\}}\, \mathcal{GIG}\left(\frac{qN_{0g} + 1}{2},\, \mathrm{tr}(D_{\nu_g}^{-1} \underline{\Delta}_g^t\, \Omega_y^{-1} \underline{\Delta}_g),\, 2\,\gamma_g\right) + \mathbb{1}_{\{\Delta_g = 0\}}\, \Gamma(\alpha_g,\, \gamma_g)$$

where $D_{\nu_g} = \mathrm{diag}(\nu_{g1}, \ldots, \nu_{g\kappa_g})$.

– *The parameter* π *satisfies, for* $j \in [\![1, 2]\!]$,

$$\pi_j \,|\, \Theta_{\pi_j} \;\sim\; \beta\big(A_j + a_j,\, B_j + b_j\big).$$

where $A_1 = G_0$, $B_1 = m - G_0$, $A_2 = J_0$ *and* $B_2 = p - N_0$.

Proof. See Section 3.5.3. □

It only remains to give the explicit results for the particular case $q = 1$. The direct links form a row vector such that $\Delta^t \in \mathbb{R}^p$ with groups $\underline{\Delta}_g^t \in \mathbb{R}^{\kappa_g}$ $(1 \leqslant g \leqslant m)$ containing predictors $\Delta_{gi} \in \mathbb{R}$ $(1 \leqslant i \leqslant \kappa_g)$, and the precision matrix of the responses reduces to $\omega_y > 0$. According to the parametrization of the distributions (see Section 3.1), the corresponding prior distribution of ω_y is $\Gamma(\frac{u}{2}, \frac{1}{2\,v})$ for $u, v > 0$, like in the other settings, and the one of ν_{gi} is $\mathcal{E}(\ell_{gi})$ for $\ell_{gi} > 0$. The other priors are unchanged.

Corollary 3.4.2. *In the hierarchical model* (6) *with* $q = 1$*, the conditional posterior distributions are as follows.*

– *The parameter* Δ_{gi} *satisfies, for* $g \in [\![1, m]\!]$ *and* $i \in [\![1, \kappa_g]\!]$*,*

$$\Delta_{gi} \,|\, \Theta_{\Delta_{gi}} \;\sim\; (1 - p_{gi})\, \mathcal{N}(-s_{gi}\, h_{gi},\, s_{gi}\, \omega_y) + p_{gi}\, \delta_0$$

where

$$h_{gi} = \omega_y \langle X_{gi},\, Y \rangle + \sum_{h,j \neq g,i} \langle X_{gi},\, X_{hj} \rangle\, \Delta_{hj}, \qquad s_{gi} = \frac{\nu_{gi}\, \lambda_g}{1 + \nu_{gi}\, \lambda_g\, \|X_{gi}\|^2}$$

and

$$p_{gi} = \frac{\rho_{gi}}{\rho_{gi} + (1 - \pi_1)\,(1 - \pi_2)\,(1 + \nu_{gi}\, \lambda_g\, \|X_{gi}\|^2)^{-\frac{1}{2}}\, \exp\!\Big(\frac{s_{gi}\, h_{gi}^2}{2\, \omega_y}\Big)}$$

in which $\rho_{gi} = (1 - \pi_1)\, \pi_2\, \mathbb{1}_{\{\Delta_{g\setminus i} \neq 0\}} + \pi_1\, \mathbb{1}_{\{\Delta_{g\setminus i} = 0\}}$*.*

– *The parameter* ω_y *satisfies*

$$\omega_y \,|\, \Theta_{\omega_y} \;\sim\; \mathcal{GIG}\bigg(\frac{n - p + N_0 + u}{2},\, \Delta\,(X^t\, X + D_{\lambda\nu}^{-1})\,\Delta^t,\, Y^t\, Y + \frac{1}{v}\bigg)$$

where $D_{\lambda\nu} = \mathrm{diag}(\nu_{11}\, \lambda_1, \ldots, \nu_{1\kappa_1}\, \lambda_1, \ldots, \nu_{m1}\, \lambda_m, \ldots, \nu_{m\kappa_m}\, \lambda_m)$*.*

– *The parameter* ν *satisfies, for* $g \in [\![1, m]\!]$ *and* $i \in [\![1, \kappa_g]\!]$*,*

$$\nu_{gi} \,|\, \Theta_{\nu_{gi}} \;\sim\; \mathbb{1}_{\{\Delta_{gi} \neq 0\}}\, \mathcal{GIG}\bigg(\frac{1}{2},\, \frac{\Delta_{gi}^2}{\lambda_g\, \omega_y},\, 2\, \ell_{gi}\bigg) + \mathbb{1}_{\{\Delta_{gi} = 0\}}\, \mathcal{E}(\ell_{gi}).$$

– *The parameter* λ *satisfies, for* $g \in [\![1, m]\!]$*,*

$$\lambda_g \,|\, \Theta_{\lambda_g} \;\sim\; \mathbb{1}_{\{\Delta_g \neq 0\}}\, \mathcal{GIG}\bigg(\frac{N_{0g} + 1}{2},\, \frac{\underline{\Delta}_g\, D_{\nu_g}^{-1}\, \underline{\Delta}_g^t}{\omega_y},\, 2\, \gamma_g\bigg) + \mathbb{1}_{\{\Delta_g = 0\}}\, \Gamma(\alpha_g,\, \gamma_g)$$

where $D_{\nu_g} = \mathrm{diag}(\nu_{g1}, \ldots, \nu_{g\kappa_g})$*.*

– *The parameter* π *satisfies, for* $j \in [\![1, 2]\!]$*,*

$$\pi_j \,|\, \Theta_{\pi_j} \;\sim\; \beta\big(A_j + a_j,\, B_j + b_j\big).$$

where $A_1 = G_0$*,* $B_1 = m - G_0$*,* $A_2 = J_0$ *and* $B_2 = p - N_0$*.*

Proof. This is a consequence of Proposition 3.4.1. □

In the simulation study of Section 3.6.1, Scen. 5 and 6 are dedicated to the sparse-group-sparse setting. Now, let us prove our assertions by a few computational steps.

3.5 Conditional posterior distributions

This section is devoted to the proofs of our assertions for the different settings.

3.5.1 The sparse setting: proof of Proposition 3.2.1

First of all, the full posterior distribution of the parameters conditional on X and Y satisfies

$$\begin{aligned} p(\Delta, \Omega_y, \lambda, \pi \,|\, Y, X) &\propto p(Y \,|\, X, \Delta, \Omega_y)\, p(\Delta \,|\, \Omega_y, \lambda, \pi)\, p(\lambda)\, p(\Omega_y)\, p(\pi) \\ &\propto |\Omega_y|^{\frac{n}{2}} \exp\left(-\frac{1}{2} \left\| (Y + X\,\Delta^t\,\Omega_y^{-1})\,\Omega_y^{\frac{1}{2}} \right\|_F^2\right) \\ &\quad \times \prod_{i=1}^{p} \left[\frac{1-\pi}{\sqrt{\lambda_i^q\,|\Omega_y|}} \exp\left(-\frac{\Delta_i^t\,\Omega_y^{-1}\Delta_i}{2\,\lambda_i}\right) \mathbb{1}_{\{\Delta_i \neq 0\}} \right. \\ &\qquad \left. + \,\pi\,\mathbb{1}_{\{\Delta_i = 0\}} \right] \lambda_i^{\frac{1}{2}(q+1)-1}\, \mathrm{e}^{-\ell_i\,\lambda_i} \\ &\quad \times |\Omega_y|^{\frac{u-q-1}{2}} \exp\left(-\frac{\mathrm{tr}(V^{-1}\,\Omega_y)}{2}\right) \pi^{\,a-1}\,(1-\pi)^{\,b-1}. \end{aligned} \tag{9}$$

On the one hand, exploiting the cyclic property of the trace, a tedious calculation shows that, for all $1 \leqslant i \leqslant p$,

$$\begin{aligned} \left\| (Y + X\,\Delta^t\,\Omega_y^{-1})\,\Omega_y^{\frac{1}{2}} \right\|_F^2 &= \mathrm{tr}(Y^{\,t}\,Y\,\Omega_y) + 2\,\mathrm{tr}(X^t\,Y\Delta) + \mathrm{tr}(X^t\,X\,\Delta^t\,\Omega_y^{-1}\Delta) \\ &= \|X_i\|^{\,2}\,\Delta_i^t\,\Omega_y^{-1}\Delta_i + 2\sum_{j\,\neq\,i} \langle X_i,\, X_j\rangle\,\Delta_j^t\,\Omega_y^{-1}\Delta_i + 2\,X_i^{\,t}\,Y\Delta_i + T_{\neq\,i} \end{aligned} \tag{10}$$

where the term $T_{\neq\,i}$ does not depend on Δ_i. Thus,

$$\begin{aligned} p(\Delta_i \,|\, \Theta_{\Delta_i}) &\propto \exp\left(-\frac{1}{2}\,\|X_i\|^{\,2}\,\Delta_i^t\,\Omega_y^{-1}\Delta_i - \sum_{j\,\neq\,i}\langle X_i,\, X_j\rangle\,\Delta_j^t\,\Omega_y^{-1}\Delta_i - X_i^{\,t}\,Y\Delta_i\right) \\ &\quad \times \left[\frac{1-\pi}{\sqrt{\lambda_i^q\,|\Omega_y|}} \exp\left(-\frac{\Delta_i^t\,\Omega_y^{-1}\Delta_i}{2\,\lambda_i}\right)\mathbb{1}_{\{\Delta_i\,\neq\,0\}} + \pi\,\mathbb{1}_{\{\Delta_i\,=\,0\}}\right] \\ &= \exp\left(-\frac{1}{2}\,(\Delta_i + s_i\,H_i)^{\,t}\,(s_i\,\Omega_y)^{-1}\,(\Delta_i + s_i\,H_i)\right) \\ &\quad \times \exp\left(\frac{s_i\,H_i^t\,\Omega_y^{-1}\,H_i}{2}\right) \frac{1-\pi}{\sqrt{\lambda_i^q\,|\Omega_y|}}\,\mathbb{1}_{\{\Delta_i\,\neq\,0\}} + \pi\,\mathbb{1}_{\{\Delta_i\,=\,0\}} \end{aligned} \tag{11}$$

for all $1 \leqslant i \leqslant p$, where

$$H_i = \Omega_y\,Y^{\,t}\,X_i + \sum_{j\,\neq\,i}\langle X_i,\, X_j\rangle\,\Delta_j \quad \text{and} \quad s_i = \frac{\lambda_i}{1 + \lambda_i\,\|X_i\|^{\,2}}.$$

This is still a multivariate Gaussian spike-and-slab distribution such that, by renormalizing, the spike has probability

$$p_i = \mathbb{P}(\Delta_i = 0 \,|\, \Theta_{\Delta_i}) = \frac{\pi}{\pi + (1-\pi)\,(1+\lambda_i\,\|X_i\|^2)^{-\frac{q}{2}}\,\exp\!\Big(\frac{s_i\,H_i^t\,\Omega_y^{-1}H_i}{2}\Big)}.$$

On the other hand, coming back to (10), we can also write

$$\Big\|(Y + X\,\Delta^t\,\Omega_y^{-1})\,\Omega_y^{\frac{1}{2}}\Big\|_F^2 = \mathrm{tr}(Y^t\,Y\,\Omega_y) + \mathrm{tr}(\Delta\,X^t\,X\,\Delta^t\,\Omega_y^{-1}) + T_{\neq y}$$

where $T_{\neq y}$ does not depend on Ω_y. That leads, *via* (9), to

$$\begin{aligned} p(\Omega_y \,|\, \Theta_{\Omega_y}) \;&\propto\; |\Omega_y|^{\frac{n-p+N_0+u-q-1}{2}}\,\exp\!\bigg(-\frac{1}{2}\,\mathrm{tr}((Y^t\,Y + V^{-1})\,\Omega_y) \\ &\qquad -\frac{1}{2}\bigg(\mathrm{tr}(\Delta\,X^t\,X\,\Delta^t\,\Omega_y^{-1}) + \sum_{\Delta_i\neq 0}\frac{\Delta_i^t\,\Omega_y^{-1}\Delta_i}{\lambda_i}\bigg)\bigg) \\ &=\; |\Omega_y|^{\frac{n-p+N_0+u-q-1}{2}}\,\exp\!\bigg(-\frac{1}{2}\,\mathrm{tr}\Big((Y^t\,Y + V^{-1})\,\Omega_y + \Delta\,(X^t\,X + D_\lambda^{-1})\,\Delta^t\,\Omega_y^{-1}\Big)\bigg) \quad (12) \end{aligned}$$

where N_0 is given in (3) and $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Finally, it is easy to see that, for all $1 \leqslant i \leqslant p$,

$$p(\lambda_i \,|\, \Theta_{\lambda_i}) \propto \frac{1}{\sqrt{\lambda_i}}\,\exp\!\bigg(-\frac{\Delta_i^t\,\Omega_y^{-1}\Delta_i}{2\,\lambda_i} - \ell_i\,\lambda_i\bigg)\mathbb{1}_{\{\Delta_i\neq 0\}} + \lambda_i^{\frac{1}{2}(q+1)-1}\,\mathrm{e}^{-\ell_i\,\lambda_i}\,\mathbb{1}_{\{\Delta_i = 0\}} \quad (13)$$

whereas

$$p(\pi \,|\, \Theta_\pi) \propto \pi^{N_0+a-1}\,(1-\pi)^{p-N_0+b-1}. \quad (14)$$

We recognize in (11), (12), (13) and (14) the announced conditional posterior distributions, which concludes the proof. □

3.5.2 The group-sparse setting: proof of Proposition 3.3.1

The full posterior distribution of the parameters conditional on X and Y satisfies

$$\begin{aligned} p(\Delta, \Omega_y, \lambda, \pi \,|\, Y, X) \;&\propto\; p(Y \,|\, X, \Delta, \Omega_y)\,p(\Delta \,|\, \Omega_y, \lambda, \pi)\,p(\lambda)\,p(\Omega_y)\,p(\pi) \\ &\propto\; |\Omega_y|^{\frac{n}{2}}\,\exp\!\bigg(-\frac{1}{2}\,\Big\|(Y + X\,\Delta^t\,\Omega_y^{-1})\,\Omega_y^{\frac{1}{2}}\Big\|_F^2\bigg) \\ &\qquad \times \prod_{g=1}^{m}\bigg[\frac{1-\pi}{\sqrt{\lambda_g^{q\,\kappa_g}\,|\Omega_y|^{\kappa_g}}}\,\exp\!\bigg(-\frac{\mathrm{tr}(\underline{\Delta}_g^t\,\Omega_y^{-1}\underline{\Delta}_g)}{2\,\lambda_g}\bigg)\mathbb{1}_{\{\Delta_g\neq 0\}} \\ &\qquad\qquad + \pi\,\mathbb{1}_{\{\Delta_g = 0\}}\bigg]\lambda_g^{\frac{1}{2}(q\,\kappa_g+1)-1}\,\mathrm{e}^{-\ell_g\,\lambda_g} \\ &\qquad \times |\Omega_y|^{\frac{u-q-1}{2}}\,\exp\!\bigg(-\frac{\mathrm{tr}(V^{-1}\,\Omega_y)}{2}\bigg)\,\pi^{\,a-1}\,(1-\pi)^{\,b-1}. \quad (15) \end{aligned}$$

Like in the previous proof, a first important step is to note that, for all $1 \leqslant g \leqslant m$,

$$\begin{aligned}\left\|(Y + X\,\Delta^t\,\Omega_y^{-1})\,\Omega_y^{\frac{1}{2}}\right\|_F^2 &= \left\|Y\,\Omega_y^{\frac{1}{2}} + \sum_{j=1}^{m} \underline{X}_j\,\underline{\Delta}_j^t\,\Omega_y^{-\frac{1}{2}}\right\|_F^2 \\ &= \|\underline{X}_g\,\underline{\Delta}_g^t\,\Omega_y^{-\frac{1}{2}}\|_F^2 + 2\sum_{j\neq g}\operatorname{tr}(\underline{\Delta}_j\,\underline{X}_j^t\,\underline{X}_g\,\underline{\Delta}_g^t\,\Omega_y^{-1}) + 2\operatorname{tr}(\underline{X}_g^t\,Y\,\underline{\Delta}_g) + T_{\neq g} \quad (16)\end{aligned}$$

where the term $T_{\neq g}$ does not depend on $\underline{\Delta}_g$. Thus, after a tedious calculation exploiting the cyclic property of the trace, one can obtain the factorization

$$\begin{aligned}p(\underline{\Delta}_g \,|\, \Theta_{\Delta_g}) &\propto \exp\left(-\frac{1}{2}\,\|\underline{X}_g\,\underline{\Delta}_g^t\,\Omega_y^{-\frac{1}{2}}\|_F^2 - \sum_{j\neq g}\operatorname{tr}(\underline{\Delta}_j\,\underline{X}_j^t\,\underline{X}_g\,\underline{\Delta}_g^t\,\Omega_y^{-1}) - \operatorname{tr}(\underline{X}_g^t\,Y\,\underline{\Delta}_g)\right) \\ &\quad\times \left[\frac{1-\pi}{\sqrt{\lambda_g^{q\,\kappa_g}\,|\Omega_y|^{\kappa_g}}}\,\exp\left(-\frac{\operatorname{tr}(\underline{\Delta}_g^t\,\Omega_y^{-1}\underline{\Delta}_g)}{2\,\lambda_g}\right)\mathbb{1}_{\{\Delta_g\neq 0\}} + \pi\,\mathbb{1}_{\{\Delta_g = 0\}}\right] \\ &= \exp\left(-\frac{1}{2}\operatorname{tr}\big(S_g^{-1}(\underline{\Delta}_g + H_g\,S_g)^t\,\Omega_y^{-1}(\underline{\Delta}_g + H_g\,S_g)\big)\right) \\ &\quad\times \exp\left(\frac{\operatorname{tr}(H_g^t\,\Omega_y^{-1}\,H_g\,S_g)}{2}\right)\frac{1-\pi}{\sqrt{\lambda_g^{q\,\kappa_g}\,|\Omega_y|^{\kappa_g}}}\,\mathbb{1}_{\{\Delta_g\neq 0\}} + \pi\,\mathbb{1}_{\{\Delta_g = 0\}} \quad (17)\end{aligned}$$

for all $1 \leqslant g \leqslant m$, where

$$H_g = \Omega_y\,Y^t\,\underline{X}_g + \sum_{j\neq g}\underline{\Delta}_j\,\underline{X}_j^t\,\underline{X}_g \quad \text{and} \quad S_g = \lambda_g\big(I_{\kappa_g} + \lambda_g\,\underline{X}_g^t\,\underline{X}_g\big)^{-1}.$$

We recognize the announced Gaussian spike-and-slab distribution, and the probability of the spike is given, after renormalization, by

$$p_g = \mathbb{P}(\underline{\Delta}_g = 0\,|\,\Theta_{\Delta_g}) = \frac{\pi}{\pi + (1-\pi)\,|I_{\kappa_g} + \lambda_g\,\underline{X}_g^t\,\underline{X}_g|^{-\frac{q}{2}}\,\exp\Big(\frac{\operatorname{tr}(H_g^t\,\Omega_y^{-1}H_g\,S_g)}{2}\Big)}.$$

Following the same lines as the ones used to establish (12), we obtain from (15) the conditional distribution

$$\begin{aligned}p(\Omega_y \,|\, \Theta_{\Omega_y}) &\propto |\Omega_y|^{\frac{n-p+N_0+u-q-1}{2}}\,\exp\left(-\frac{1}{2}\operatorname{tr}((Y^t\,Y + V^{-1})\,\Omega_y)\right. \\ &\quad\left. -\frac{1}{2}\left(\operatorname{tr}(\Delta\,X^t\,X\,\Delta^t\,\Omega_y^{-1}) + \sum_{\Delta_g\neq 0}\frac{\operatorname{tr}(\underline{\Delta}_g^t\,\Omega_y^{-1}\underline{\Delta}_g)}{\lambda_g}\right)\right) \\ &= |\Omega_y|^{\frac{n-p+N_0+u-q-1}{2}}\,\exp\left(-\frac{1}{2}\operatorname{tr}\big((Y^t\,Y + V^{-1})\,\Omega_y + \Delta\,(X^t\,X + D_\lambda^{-1})\,\Delta^t\,\Omega_y^{-1}\big)\right) \quad (18)\end{aligned}$$

where $D_\lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_1,\ldots,\lambda_m,\ldots,\lambda_m)$ with each λ_g duplicated κ_g times, and since we can note

that, due to the continuous nature of $\Delta \,|\{\Delta \neq 0\}$,

$$\sum_{g=1}^{m} \kappa_g \mathbb{1}_{\{\Delta_g \neq 0\}} = p - N_0$$

for N_0 given in (3). Next, we obtain in a simpler way that, for all $1 \leqslant g \leqslant m$,

$$p(\lambda_g \,|\, \Theta_{\lambda_g}) \quad \propto \quad \frac{1}{\sqrt{\lambda_g}} \exp\left(-\frac{\mathrm{tr}(\underline{\Delta}_g^t \, \Omega_y^{-1} \underline{\Delta}_g)}{2\,\lambda_g} - \ell_g \, \lambda_g\right) \mathbb{1}_{\{\Delta_g \neq 0\}} + \lambda_g^{\frac{1}{2}(q\,\kappa_g+1)-1} \, \mathrm{e}^{-\ell_g \, \lambda_g} \, \mathbb{1}_{\{\Delta_g = 0\}}. \quad (19)$$

Finally,

$$p(\pi \,|\, \Theta_\pi) \propto \pi^{G_0+a-1} \, (1-\pi)^{m-G_0+b-1} \quad (20)$$

where G_0 is defined in (5). We can check that the conditional distributions (17), (18), (19) and (20) correspond to the ones announced in the proposition, which concludes the proof. □

3.5.3 The sparse-group-sparse setting: proof of Proposition 3.4.1

The full posterior distribution of the parameters conditional on X and Y satisfies

$$\begin{aligned}
p(\Delta, \Omega_y, \nu, \lambda, \pi \,|\, Y, X) \quad &\propto \quad p(Y \,|\, X, \Delta, \Omega_y) \, p(\Delta \,|\, \Omega_y, \nu, \lambda, \pi) \, p(\nu) \, p(\lambda) \, p(\Omega_y) \, p(\pi) \\
&\propto \quad |\Omega_y|^{\frac{n}{2}} \exp\left(-\frac{1}{2} \left\| (Y + X\,\Delta^t\,\Omega_y^{-1})\,\Omega_y^{\frac{1}{2}} \right\|_F^2\right) \\
&\qquad \times \prod_{g=1}^{m} \Bigg[\left((1-\pi_1)\, P_g \, \mathbb{1}_{\{\Delta_g \neq 0\}} + \pi_1 \, \mathbb{1}_{\{\Delta_g = 0\}}\right) \\
&\qquad\qquad \times \lambda_g^{\frac{1}{2}(q\,\kappa_g+1)-1} \, \mathrm{e}^{-\gamma_g \, \lambda_g} \prod_{i=1}^{\kappa_g} \nu_{gi}^{\frac{1}{2}(q+1)-1} \, \mathrm{e}^{-\ell_{gi} \, \nu_{gi}} \Bigg] \\
&\qquad \times |\Omega_y|^{\frac{u-q-1}{2}} \exp\left(-\frac{\mathrm{tr}(V^{-1}\,\Omega_y)}{2}\right) \prod_{j=1}^{2} \pi_j^{\,a_j-1} \, (1-\pi_j)^{\,b_j-1} \qquad (21)
\end{aligned}$$

where, for $1 \leqslant g \leqslant m$,

$$P_g = \prod_{i=1}^{\kappa_g} \left[\frac{1-\pi_2}{\sqrt{(\nu_{gi}\,\lambda_g)^{\,q}\,|\Omega_y|}} \exp\left(-\frac{\Delta_{gi}^t \, \Omega_y^{-1} \Delta_{gi}}{2\,\nu_{gi}\,\lambda_g}\right) \mathbb{1}_{\{\Delta_{gi} \neq 0\}} + \pi_2 \, \mathbb{1}_{\{\Delta_{gi} = 0\}} \right].$$

Using the same decompositions as (10) or (16), the full posterior distribution given above leads to

$$\begin{aligned} p(\Delta_{gi} \,|\, \Theta_{\Delta_{gi}}) \;\; &\propto \;\; \exp\!\left(-\frac{1}{2}\,\|X_{gi}\|^{\,2}\,\Delta_{gi}^{t}\,\Omega_{y}^{-1}\Delta_{gi} - \sum_{h,j\,\neq\,g,i}\langle X_{gi},\,X_{hj}\rangle\,\Delta_{hj}^{t}\,\Omega_{y}^{-1}\Delta_{gi} - X_{gi}^{t}\,Y\Delta_{gi}\right) \\ &\qquad \times \left[(1-\pi_1)\left[\frac{1-\pi_2}{\sqrt{(\nu_{gi}\,\lambda_g)^{\,q}\,|\Omega_y|}}\;\exp\!\left(-\frac{\Delta_{gi}^{t}\,\Omega_{y}^{-1}\Delta_{gi}}{2\,\nu_{gi}\,\lambda_g}\right)\mathbb{1}_{\{\Delta_{gi}\,\neq\,0\}}\right.\right. \\ &\qquad\qquad \left.\left. +\;\pi_2\;\mathbb{1}_{\{\Delta_{gi}\,=\,0\}}\right]\mathbb{1}_{\{\Delta_{g}\,\neq\,0\}} + \pi_1\;\mathbb{1}_{\{\Delta_{g}\,=\,0\}}\right] \\ &= \;\; \exp\!\left(-\frac{1}{2}\,(\Delta_{gi}+s_{gi}\,H_{gi})^{\,t}\,(s_{gi}\,\Omega_y)^{-1}\,(\Delta_{gi}+s_{gi}\,H_{gi})\right) \\ &\qquad \times\;\exp\!\left(\frac{s_{gi}\,H_{gi}^{t}\,\Omega_{y}^{-1}\,H_{gi}}{2}\right)\frac{(1-\pi_1)\,(1-\pi_2)}{\sqrt{(\nu_{gi}\,\lambda_g)^{\,q}\,|\Omega_y|}}\;\mathbb{1}_{\{\Delta_{gi}\,\neq\,0\}} \\ &\qquad +\;\big((1-\pi_1)\,\pi_2\;\mathbb{1}_{\{\Delta_{g\setminus i}\,\neq\,0\}}\;+\pi_1\;\mathbb{1}_{\{\Delta_{g\setminus i}\,=\,0\}}\big)\;\mathbb{1}_{\{\Delta_{gi}\,=\,0\}} \end{aligned} \tag{22}$$

for $1 \leqslant g \leqslant m$ and $1 \leqslant i \leqslant \kappa_g$, where $\underline{\Delta}_{g\setminus i}$ is $\underline{\Delta}_g$ deprived of Δ_{gi},

$$H_{gi} = \Omega_y\,Y^{\,t}\,X_{gi} + \sum_{h,j\,\neq\,g,i}\langle X_{gi},\,X_{hj}\rangle\,\Delta_{hj} \qquad \text{and} \qquad s_{gi} = \frac{\nu_{gi}\,\lambda_g}{1+\nu_{gi}\,\lambda_g\,\|X_{gi}\|^{\,2}}.$$

Here, we used the binary equalities stemming from $\{\Delta_{gi} \neq 0\} \cap \{\underline{\Delta}_g \neq 0\} = \{\Delta_{gi} \neq 0\}$, $\{\Delta_{gi} = 0\} \cap \{\underline{\Delta}_g \neq 0\} = \{\Delta_{gi} = 0\} \cap \{\underline{\Delta}_{g\setminus i} \neq 0\}$ and $\{\Delta_{gi} = 0\} \cap \{\underline{\Delta}_g = 0\} = \{\Delta_{gi} = 0\} \cap \{\underline{\Delta}_{g\setminus i} = 0\}$, which turn out to be very useful to separate Δ_{gi} and $\Theta_{\Delta_{gi}}$. This is characteristic of a multivariate Gaussian spike-and-slab distribution. By renormalizing, one can see that the spike has probability

$$p_{gi} = \mathbb{P}(\Delta_{gi} = 0 \,|\, \Theta_{\Delta_{gi}}) = \frac{\rho_{gi}}{\rho_{gi} + (1-\pi_1)\,(1-\pi_2)\,(1+\nu_{gi}\,\lambda_g\,\|X_{gi}\|^{\,2})^{-\frac{q}{2}}\;\exp\!\Big(\frac{s_{gi}\,H_{gi}^{t}\,\Omega_{y}^{-1}\,H_{gi}}{2}\Big)}$$

with

$$\rho_{gi} = (1-\pi_1)\,\pi_2\;\mathbb{1}_{\{\Delta_{g\setminus i}\,\neq\,0\}}\;+\pi_1\;\mathbb{1}_{\{\Delta_{g\setminus i}\,=\,0\}}.$$

Next, following (21) and the reasoning used to establish (12), we may also write

$$\begin{aligned} p(\Omega_y \,|\, \Theta_{\Omega_y}) \;\; &\propto \;\; |\Omega_y|^{\frac{n-p+N_0+u-q-1}{2}}\;\exp\!\left(-\frac{1}{2}\,\mathrm{tr}((Y^{\,t}\,Y+V^{-1})\,\Omega_y)\right. \\ &\qquad \left. -\,\frac{1}{2}\left(\mathrm{tr}(\Delta\,X^{t}\,X\,\Delta^{\,t}\,\Omega_{y}^{-1}) + \sum_{\Delta_{gi}\,\neq\,0}\frac{\Delta_{gi}^{t}\,\Omega_{y}^{-1}\Delta_{gi}}{\nu_{gi}\,\lambda_g}\right)\right) \\ &= \;\; |\Omega_y|^{\frac{n-p+N_0+u-q-1}{2}}\;\exp\!\left(-\frac{1}{2}\,\mathrm{tr}\big((Y^{\,t}\,Y+V^{-1})\,\Omega_y + \Delta\,(X^{t}\,X+D_{\lambda\nu}^{-1})\,\Delta^{\,t}\,\Omega_{y}^{-1}\big)\right) \end{aligned} \tag{23}$$

where N_0 is given in (3) and $D_{\lambda\nu} = \mathrm{diag}(\nu_{11}\lambda_1, \ldots, \nu_{1\kappa_1}\lambda_1, \ldots, \nu_{m1}\lambda_m, \ldots, \nu_{m\kappa_m}\lambda_m)$. The shrinkage parameters ν and λ are easier to handle. For $1 \leqslant g \leqslant m$ and $1 \leqslant i \leqslant \kappa_g$,

$$\begin{aligned} p(\nu_{gi} \,|\, \Theta_{\nu_{gi}}) \quad &\propto \quad \frac{1}{\sqrt{\nu_{gi}}} \exp\left(-\frac{\Delta_{gi}^t\, \Omega_y^{-1} \Delta_{gi}}{2\,\nu_{gi}\,\lambda_g} - \ell_{gi}\,\nu_{gi}\right) \mathbb{1}_{\{\Delta_{gi} \neq 0\}} \\ &\quad + \nu_{gi}^{\frac{1}{2}(q+1)-1}\, \mathrm{e}^{-\ell_{gi}\,\nu_{gi}}\, \mathbb{1}_{\{\Delta_{gi} = 0\}} \end{aligned} \tag{24}$$

whereas

$$\begin{aligned} p(\lambda_g \,|\, \Theta_{\lambda_g}) \quad &\propto \quad \lambda_g^{\frac{qN_{0g}-1}{2}} \exp\left(-\frac{\mathrm{tr}(D_{\nu_g}^{-1} \Delta_g^t\, \Omega_y^{-1} \Delta_g)}{2\,\lambda_g} - \gamma_g\,\lambda_g\right) \mathbb{1}_{\{\Delta_g \neq 0\}} \\ &\quad + \lambda_g^{\frac{1}{2}(q\,\kappa_g+1)-1}\, \mathrm{e}^{-\gamma_g\,\lambda_g}\, \mathbb{1}_{\{\Delta_g = 0\}} \end{aligned} \tag{25}$$

where N_{0g} is defined in (7) and $D_{\nu_g} = \mathrm{diag}(\nu_{g1}, \ldots, \nu_{g\kappa_g})$. Finally,

$$p(\pi_1 \,|\, \Theta_{\pi_1}) \propto \pi_1^{G_0+a_1-1}\,(1-\pi_1)^{m-G_0+b_1-1} \tag{26}$$

and

$$p(\pi_2 \,|\, \Theta_{\pi_2}) \propto \pi_2^{J_0+a_2-1}\,(1-\pi_2)^{p-N_0+b_2-1} \tag{27}$$

where G_0 and J_0 are given in (5) and (8), respectively. For the latter result, we used the fact that the number of non-zero columns in the non-zero groups must coincide with the number of non-zero columns of Δ, that is $p - N_0$. Like in the previous proofs, we recognize the announced conditional distributions in (22), (23), (24), (25), (26) and (27). That concludes these tedious calculations. □

3.5.4 Proof of Proposition 3.3.3

The result is obtained by following the steps of the proof of Theorem 2.1 in [YN20] but, beforehand, we need to clarify a few points to extend the reasoning of the authors from $q = 1$ to $q \geqslant 1$ and take into account the adaptative shrinkage. For any model k, let $\mathcal{K} = \{k \text{ is selected}\}$ so that $\mathcal{K} = \mathcal{T}$ when the true model t is considered. First, recall that λ and π are fixed and rewrite (15) like

$$\begin{aligned} \mathbb{P}_\Delta(\mathcal{K} \,|\, Y, X, \Omega_y) \quad &\propto \quad \exp\left(-\frac{1}{2}\left\|(Y + X_{(k)}\,\Delta_{(k)}^t\,\Omega_y^{-1})\,\Omega_y^{\frac{1}{2}}\right\|_F^2\right) \\ &\qquad \times \frac{(1-\pi)^{|k|}}{\pi^{|k|}\,\sqrt{|\Lambda_k|^q\,|\Omega_y|^{k_r}}} \exp\left(-\frac{\mathrm{tr}(\Delta_{(k)}^t\,\Omega_y^{-1}\Delta_{(k)}\,D_k^{-1})}{2}\right) \\ &\propto \quad \frac{(1-\pi)^{|k|}}{\pi^{|k|}\,\sqrt{|\Lambda_k|^q\,|\Omega_y|^{k_r}}} \exp\left(\frac{\mathrm{tr}(\widetilde{\Delta}_{(k)}\,F_k\,\widetilde{\Delta}_{(k)}^t\,\Omega_y^{-1})}{2}\right) \\ &\qquad \times \exp\left(-\frac{1}{2}\,\mathrm{tr}\Big((\Delta_{(k)} - \widetilde{\Delta}_{(k)})\,F_k\,(\Delta_{(k)} - \widetilde{\Delta}_{(k)})^t\,\Omega_y^{-1}\Big)\right) \end{aligned} \tag{28}$$

where $F_k = D_k^{-1} + X_{(k)}^t X_{(k)}$, $D_k = \mathrm{diag}((\lambda_\ell, \dots, \lambda_\ell)_{\ell \in k})$ with each λ_ℓ duplicated κ_ℓ times, $k_r = \|(\kappa_\ell)_{\ell \in k}\|_1$, $\Lambda_k = \mathrm{diag}((\lambda_\ell^{\kappa_\ell})_{\ell \in k})$ and

$$\widetilde{\Delta}_{(k)} = -\Omega_y \, Y^t \, X_{(k)} \, F_k^{-1}.$$

Then, integrating over $\Delta_{(k)}$, it follows (see Def. 3.1.1 with $\Sigma_1 = \Omega_y$ and $\Sigma_2 = F_k^{-1}$) that

$$\begin{aligned}
\mathbb{P}(\mathcal{K} \,|\, Y, X, \Omega_y) &= \int_{\mathbb{R}^{q \times \kappa_k}} \mathbb{P}_\Delta(\mathcal{K} \,|\, Y, X, \Omega_y, \lambda, \pi) \, \mathrm{d}\Delta_{(k)} \\
&\propto \frac{(1-\pi)^{|k|}}{\pi^{|k|} \sqrt{|\Lambda_k|^q \, |F_k|^q}} \exp\left(\frac{\mathrm{tr}(\widetilde{\Delta}_{(k)} \, F_k \, \widetilde{\Delta}_{(k)}^t \, \Omega_y^{-1})}{2} \right) \\
&\propto \left(\frac{1-\pi}{\pi} \right)^{|k|} |\Lambda_k|^{-\frac{q}{2}} \, |F_k|^{-\frac{q}{2}} \, \exp\left(-\frac{1}{2} \, \mathrm{tr}\big(Y^{*\,t} \, (I_n - X_{(k)} \, F_k^{-1} \, X_{(k)}^t) \, Y^* \big) \right) \\
&= \left(\frac{1-\pi}{\pi} \right)^{|k|} |\Lambda_k|^{-\frac{q}{2}} \, |F_k|^{-\frac{q}{2}} \, \exp\left(-\frac{1}{2} \, \left(\mathrm{RSS}_k(\widetilde{\Delta}_{(k)}^*) + \big\| \widetilde{\Delta}_{(k)}^* \, D_k^{-\frac{1}{2}} \big\|_F^2 \right) \right)
\end{aligned}$$

where $Y^* = Y \, \Omega_y^{\frac{1}{2}}$, $\widetilde{\Delta}_{(k)}^* = \Omega_y^{-\frac{1}{2}} \, \widetilde{\Delta}_{(k)}$ and $\mathrm{RSS}_k : H \in \mathbb{R}^{q \times \kappa_k} \mapsto \|Y^* - X_{(k)} \, H^t\|_F^2$ is the residual sum of squares function in the renormalized linear model indexed by k, that is

$$Y^* = -X_{(k)} \, \Delta_{(k)}^t \, \Omega_y^{-\frac{1}{2}} + E^*$$

with $E^* = E \, \Omega_y^{\frac{1}{2}} \sim \mathcal{MN}_{n \times q}(0, \, I_n, \, I_q)$. Thus, the so-called posterior ratio between any false model k and t is given by

$$\mathrm{PR}(k, t) = \frac{\mathbb{P}(\mathcal{K} \,|\, Y, X, \Omega_y)}{\mathbb{P}(\mathcal{T} \,|\, Y, X, \Omega_y)} = \frac{Q_k}{Q_t} \left(\frac{1-\pi}{\pi} \right)^{|k|-|t|} \mathrm{e}^{-\frac{1}{2}(\widetilde{R}_k - \widetilde{R}_t)}$$

with $Q_k = |\Lambda_k|^{-\frac{q}{2}} \, |F_k|^{-\frac{q}{2}}$ and $\widetilde{R}_k = \mathrm{RSS}_k(\widetilde{\Delta}_{(k)}^*) + \|\widetilde{\Delta}_{(k)}^* \, D_k^{-\frac{1}{2}}\|_F^2$, using the notation of [YN20]. In particular, due to the generalized ridge penalty,

$$\widetilde{\Delta}_{(k)}^* = \arg\min_H \, \left(\mathrm{RSS}_k(H) + \big\| H D_k^{-\frac{1}{2}} \big\|_F^2 \right) \tag{29}$$

so that for nested models k_1 and k_2 (with $k_1 \subseteq k_2$), we must have $\widetilde{R}_{k_2} \leqslant \widetilde{R}_{k_1}$. Let also $R_k = \|(I_n - \Pi_{(k)}) \, Y^*\|_F^2 = \|(I_q \otimes (I_n - \Pi_{(k)})) \, \mathrm{vec}(Y^*)\|_2^2$. Cochran's theorem entails the chi-squared distributions $R_t \sim \chi^2(q \, (n - r_t))$ and $R_t - R_k \sim \chi^2(q \, (r_k - r_t))$ for any 'bigger' model $k \supset t$ and $q \geqslant 1$. Combining all these preliminary considerations, the strategy of [YN20] now applies and leads, under our revised hypotheses, to

$$\frac{1 - \mathbb{P}(\mathcal{T} \,|\, Y, X, \Omega_y)}{\mathbb{P}(\mathcal{T} \,|\, Y, X, \Omega_y)} = \sum_{k \neq t} \mathrm{PR}(k, t) \xrightarrow{\mathbb{P}} 0.$$

□

3.6 Empirical results

In this section, let us call (s), (gs) and (sgs) the related settings, and let us denote by (ad) the adaptative shrinkage and by (gl) the global shrinkage. First of all, these models contain many hyperparameters that have to be carefully tuned. Our experiments showed that, unsurprinsingly, the results are strongly impacted by the prior amount of shrinkage on Δ, driven by ℓ and even by γ for (sgs). Apart from the usual cross-validation procedures, we could stay in line with our Bayesian approach and suggest conjugate Gamma hyperpriors. This is very easy to implement, but the hyperparameters are now replaced by other hyperparameters and the same questions arise. Instead, like in [XG15] and [LMPS17], we follow the idea of [PC08] and we use a Monte-Carlo EM algorithm. By way of example, from the full posterior probability (9) and since $\lambda_i \sim \Gamma(\alpha,\, \ell_i)$ for all i, it is not hard to see that, with (s),

$$\ln p(\Delta, \Omega_y, \lambda, \pi \,|\, Y, X) = \sum_{i=1}^{p} (\alpha \, \ln \ell_i - \ell_i \, \lambda_i) + T_{\neq \ell}$$

where the term $T_{\neq \ell}$ does not depend on ℓ. Thus, the k-th iteration of the EM algorithm should lead to

$$\ell_i^{(k)} = \frac{\frac{1}{2}(q+1)}{\mathbb{E}^{(k-1)}[\lambda_i \,|\, Y,\, X]} \quad \text{and} \quad \ell^{\,(k)} = \frac{\frac{p}{2}(q+1)}{\sum_{i=1}^{p} \mathbb{E}^{(k-1)}[\lambda_i \,|\, Y,\, X]}$$

for the adaptative shrinkage and the global shrinkage ($\lambda_i = \lambda$), respectively. The intractable conditional expectations are then estimated with the help of the Gibbs samples. For (gs), the results are mainly the same as above (replace $q+1$ by $q\kappa_g + 1$ in the first case, $p(q+1)$ by $qp+m$ in the second case and consider $1 \leqslant g \leqslant m$ instead of $1 \leqslant i \leqslant p$), and similar results also follow with (sgs). Recall that our definitions of the adaptative and global shrinkages are given in the corresponding sections, in the description of the hierarchical models. The tuning of u and V (or v) is actually trickier. Because $\mathbb{E}[\Omega_y] = uV$, we set $V = \frac{1}{u} I_q$ and u is conveniently chosen to be the smallest integer such that Ω_y is (almost surely) invertible, that is $u = q$ (see *e.g.* [BVF98]). This is particularly adapted when the dataset is standardized. Finally, a and b reflect the degree of sparsity to introduce in the direct links. We can set $a \gg b$ to promote sparse settings, which is potentially interesting when $p \gg n$, but $a = b = 1$ is a standard non-informative choice and $a < b$ may also be useful for variable selection (see *e.g.* the real dataset of Section 3.6.2). They can be chosen from a cross-validation step (for prediction purposes) or to enforce some degree of sparsity (for selection purposes), just like a practitioner manages the tuning parameter of the Lasso. The posterior median is used to estimate Δ and get sparsity whereas the posterior mean is used to estimate Ω_y. Indeed, we don't want to impose any sparsity on Ω_y (q is small), so we decided to retain this standard choice. But the concern is much greater for Δ because some coordinates must be exactly zero. This is the reason why the posterior median seemed a more appropriate choice (in particular, it suffices for the sampler to generate zeros more than half the time for the empirical posterior median to be zero). Due to the huge amount of calculations in the simulations, the estimations are made on the basis of 3000 iterations of the sampler in which the first 2000 are burn-ins. This is revised upwards for the real data (10000 iterations with 5000 burn-ins).

Remark 3.6.1. To the best of our knowledge, there is no simple way to sample from the $\mathcal{MGIG}_d$ distribution as soon as $d > 1$. The recent method described in Section 3.3.2 of [FKS20], relying on the

Matsumoto-Yor property (see Theorem 3.1 of [MW06]) to get a $\mathcal{MGIG}_d$ sample from the very standard $\mathcal{GIG}$ and $\mathcal{W}_d$ distributions, is unfortunately inapplicable in our context. Indeed, for example in the sparse setting, that would require finding $z \in \mathbb{R}^q$ such that $Y^t Y + V^{-1} = b\, zz^t$ for some $b > 0$, which is clearly impossible since $Y^t Y + V^{-1}$ has full rank. In [FB16], the authors show that $\mathcal{MGIG}_d(\nu, A, B)$ is a unimodal distribution of which mode $M \in \mathbb{S}_{++}^d$ is the unique solution of the algebraic Riccati equation $(d+1-2\,\nu)\, M + MBM = A$, and a standard importance sampling approach follows for the mean of the distribution. Our fallback solution is to solve this Riccati equation at each step and to replace all $\mathcal{MGIG}_d$ random variables by the (unique) mode of the consecutive distributions. To assess the credibility of this *ad hoc* sampling, the 'oracle' models in which Ω_y and the shrinkage parameters are known are added to the simulations. We will see that, despite an unavoidable loss, the results remain pretty consistent. In particular, the support recovery does not appear to be impacted.

3.6.1 A simulation study

In this empirical section, the matrix of order $d \geqslant 1$ given by

$$C_d = \left(\rho^{|i-j|}\right)_{1 \leqslant i,j \leqslant d}$$

will be used as a typical covariance structure, for some $0 \leqslant \rho < 1$. Thus, the precision matrices will be chosen as a multiple of C_d^{-1} to keep the same guideline in our simulations. The responses

$$Y_i = B^t X_i + E_i$$

are generated through relations (1) where, for all $1 \leqslant k \leqslant n$, $E_i \sim \mathcal{N}(0, R)$. Because our models assume prior independence (or group-independence) in the columns of Δ, it seems necessary to look at the influence of correlation among the predictors. So the standard choice $X_k \sim \mathcal{N}(0, I_p)$ is first considered, but in some cases we will also test $X_k \sim \mathcal{N}(0, C_p)$ for $\rho = 0.5$ and $\rho = 0.9$ to introduce a significant correlation between close predictors (see Figure 3.1). For each experiment, the support recovery of Δ is evaluated thanks to the so-called F-score given by

$$F = \frac{2\, p_r\, r_e}{p_r + r_e} \quad \text{where} \quad p_r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad \text{and} \quad r_e = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

are the precision and the recall, respectively, and where T/F and P/N stand for true/false and positive/negative. To assess prediction skills, n_e randomly chosen observations are used for estimation (for different n_e) and the remaining $n_v = n - n_e = 100$ independent observations serve to compute the mean squared prediction error (MSPE). The results are compared to the ones obtained *via* the penalized maximum of likelihood (PML) approach of [YZ14] thanks to the correctly adapted implementations of [CMHR17] and [OOJP21], with a cross-validated tuning parameter. In addition, we compute the sparse precision matrix estimations given by the graphical Lasso (GLasso) of [FHT08], and by the CLIME algorithm of [CLL11], using the `R` packages `glasso` and `fastclime`, respectively. Note that we always keep a small value for q, so Δ is penalized but not Ω_y when possible (PML and GLasso). Finally, the recent approach of [RSZZ15], called ANT and based on the individual estimations of the partial correlations, is

also implemented. Unlike PML, GLasso and CLIME, sparsity is not the result of penalizations for ANT but, instead, a threshold is deduced from the asymptotic normality of the estimates to decide which are significant and which can be set to zero. Let us add some preliminary comments about the methods compared in these simulations, all related to high-dimensional precision matrix estimation.

– There is a important advantage in favor of our Bayesian approaches, PML and ANT because they do not need the estimation of $\Omega_x \in \mathbb{S}_{++}^p$. Indeed, extracting the estimation of $\Delta \in \mathbb{R}^{q\times p}$ and $\Omega_y \in \mathbb{S}_{++}^q$ from that of the full precision matrix $\Omega \in \mathbb{S}_{++}^{q+p}$ may generate a drastic bias when $p \gg q$, and that explains in particular why GLasso and CLIME give pretty bad results in what follows.

– In its standard version, ANT is not designed to produce column-sparsity or group-sparsity in Δ. So, by considering multiple testing at the column or even group level, we allow groups of coefficients to be zeroed simultaneously. We have observed that this modified ANT method (called ANT* in the simulations) loses a bit in prediction quality but is greatly improved for support recovery.

– Unfortunately, this is not appropriate for PML, GLasso and CLIME. It is therefore not surprising that they are largely outperformed by our Bayesian models and ANT* for (gs) and (sgs). Using group-penalties, which to the best of our knowledge still does not exist, should improve the results of these methods to some extent.

The seven scenarios below, from Scen. 0 to Scen. 6, as heterogeneous as possible, represent the diversity of the situations (high-dimensionality, kind of sparsity, dimension of the responses, coefficients hard to detect, etc.). We repeat each one $N = 100$ or $N = 50$ times, depending on the computation times involved, and the numerical results for $n_e = 400$ and uncorrelated predictors are summarized in Table 3.1. In addition, the evolution of MSPE is represented on Figure 3.1 for Scen. 1, 3 and 5, when n_e grows from 100 to 500, both for uncorrelated and correlated predictors. The three configurations (s), (gs) and (sgs) are tested on the grouped scenarios (from Scen. 3 to Scen. 6) with the adaptative shrinkage.

– *Scenario 0 (small dimension, no sparsity).* Let $q = 1$, $p = 5$ and set $\omega_y = 1$. We fill Δ with $\mathcal{N}(0, 2\,\omega_y)$ coefficients.

– *Scenario 1 (sparse direct links, univariate responses).* Let $q = 1$, $p = 50$ and set $\omega_y = 1$. We randomly choose 10 locations of Δ filled with $\mathcal{N}(0, \omega_y)$ coefficients while the others are zero.

– *Scenario 2 (sparse direct links, multivariate responses).* Let $q = 2$, $p = 80$ and set $\Omega_y = 2\,C_2^{-1}$ with $\rho = 0.5$. We randomly choose 10 columns of Δ filled with $\mathcal{N}_2(0, \Omega_y)$ coefficients while the others are zero.

– *Scenario 3 (group-sparse direct links, univariate responses).* Let $q = 1$, $p = 320$ and set $\omega_y = 1$. We consider $m = 5$ groups of size 100, 10, 100, 10 and 100. The two groups of size 10 are filled with $\mathcal{N}(0, 0.5\,\omega_y)$ and $\mathcal{N}(0, \omega_y)$ coefficients, respectively, while the other groups are zero.

– *Scenario 4 (group-sparse direct links, multivariate responses).* Let $q = 3$, $p = 500$ and set $\Omega_y = 3\,C_3^{-1}$ with $\rho = 0.5$. We divide the columns of Δ into $m = 25$ groups of size 20. We randomly choose 3 groups filled with $\mathcal{N}_3(0, 0.5\,\Omega_y)$, $\mathcal{N}_3(0, \Omega_y)$ and $\mathcal{N}_3(0, 1.5\,\Omega_y)$ coefficients, respectively, while the other groups are zero.

– *Scenario 5 (sparse-group-sparse direct links, univariate responses).* Let $q = 1$, $p = 150$ and set $\omega_y = 1$. We consider $m = 3$ groups of size 50. Only the second group is non-zero, into which we randomly fill 10 locations with $\mathcal{N}(0, \omega_y)$ coefficients.

– *Scenario 6 (sparse-group-sparse direct links, multivariate responses).* Let $q = 5$, $p = 1000$ and set $\Omega_y = 5\, C_5^{-1}$ with $\rho = 0.5$. We divide the columns of Δ into $m = 20$ groups of size 50, and a randomly chosen one is half filled with $\mathcal{N}_5(0, \Omega_y)$ coefficients. The others columns of Δ are zero.

Mod.	**Shr.**	**Scenario 0**			
		MSPE	F	p_r	r_e
(s-or)	-	1.01 (0.11)	<u>1.00</u>	1.00	1.00
(s)	(ad)	1.03 (0.13)	<u>1.00</u>	1.00	1.00
(s)	(gl)	1.03 (0.13)	<u>1.00</u>	1.00	1.00
PML	-	1.01 (0.16)	<u>1.00</u>	1.00	1.00
GLasso	-	<u>1.00</u> (0.15)	<u>1.00</u>	1.00	1.00
CLIME	-	<u>1.00</u> (0.15)	<u>1.00</u>	1.00	1.00
ANT*	-	1.04 (0.13)	<u>1.00</u>	1.00	1.00
Hyperparam.		$\pi = 0$			

Mod.	**Shr.**	**Scenario 1**				**Scenario 2**			
		MSPE	F	p_r	r_e	MSPE	F	p_r	r_e
(s-or)	-	<u>1.02</u> (0.13)	<u>0.95</u>	1.00	0.90	<u>0.52</u> (0.09)	<u>0.95</u>	1.00	0.90
(s)	(ad)	1.04 (0.13)	<u>0.95</u>	1.00	0.90	0.54 (0.09)	<u>0.95</u>	1.00	0.90
(s)	(gl)	1.03 (0.13)	<u>0.95</u>	1.00	0.90	0.55 (0.08)	<u>0.95</u>	1.00	0.90
PML	-	1.08 (0.15)	0.82	0.69	1.00	0.77 (0.15)	0.86	1.00	0.75
GLasso	-	2.37 (0.96)	0.78	0.77	0.80	1.74 (0.49)	0.72	0.91	0.60
CLIME	-	2.52 (0.98)	0.79	0.78	0.80	1.11 (0.35)	0.73	0.76	0.70
ANT*	-	1.25 (0.22)	0.87	0.85	0.90	1.04 (0.44)	0.90	0.89	0.91
Hyperparam.		$(25, 1)$				$(80, 1)$			

Now, let us try to summarize our observations. In terms of support recovery, the Bayesian spike-and-slab framework and the modified ANT* method give results incomparably better than the sparsity-inducing penalized approaches (PML, GLasso and CLIME). As suggested in Remark 3.3 of [OOJP21], this may be a consequence of the fact that the cross-validation steps calibrate the models to reach the best prediction error, sometimes at the cost of support recovery by picking a small penalty level. The superiority of ANT over GLasso and CLIME is recognized and discussed in [RSZZ15], but this also highlights the ability of our Bayesian models to reach good results both in prediction and in support recovery. It can also be seen that (s) gives weaker results than (sgs) in the grouped scenarios, probably

Mod.	Shr.	Scenario 3				Scenario 4			
		MSPE	F	p_r	r_e	MSPE	F	p_r	r_e
(gs-or)	-	1.03 (0.27)	1.00	1.00	1.00	0.40 (0.14)	1.00	1.00	1.00
(gs)	(ad)	1.04 (0.27)	1.00	1.00	1.00	0.45 (0.16)	1.00	1.00	1.00
(gs)	(gl)	1.04 (0.34)	1.00	1.00	1.00	0.46 (0.17)	1.00	1.00	1.00
(s)	(ad)	1.16 (0.27)	0.92	1.00	0.85	0.52 (0.18)	0.98	1.00	0.96
(sgs)	(ad)	1.07 (0.25)	0.92	1.00	0.86	0.48 (0.17)	0.99	1.00	0.98
PML	-	1.80 (0.36)	0.89	1.00	0.80	3.18 (0.53)	0.75	0.94	0.62
GLasso	-	4.23 (1.61)	0.58	0.50	0.70	9.46 (1.38)	0.46	0.66	0.35
CLIME	-	2.98 (1.22)	0.68	0.90	0.55	8.32 (1.51)	0.48	0.45	0.52
ANT*	-	1.52 (0.95)	1.00	1.00	1.00	6.53 (1.22)	1.00	1.00	1.00
Hyperparam.		$(100, 1)$ - $(5, 1)$ - $(5, 1, 25, 1)$				$(100, 1)$ - $(25, 1)$ - $(50, 1, 50, 1)$			

Mod.	Shr.	Scenario 5				Scenario 6			
		MSPE	F	p_r	r_e	MSPE	F	p_r	r_e
(sgs-or)	-	1.00 (0.15)	0.96	1.00	0.92	0.21 (0.13)	1.00	1.00	1.00
(sgs)	(ad)	1.04 (0.16)	0.95	1.00	0.91	0.24 (0.32)	1.00	1.00	1.00
(sgs)	(gl)	1.03 (0.16)	0.91	1.00	0.84	0.24 (0.33)	1.00	1.00	1.00
(s)	(ad)	1.08 (0.14)	0.93	1.00	0.87	0.29 (0.26)	0.98	1.00	0.96
(gs)	(ad)	1.24 (0.19)	0.33	0.20	1.00	0.31 (0.30)	0.67	0.50	1.00
PML	-	1.92 (0.60)	0.89	1.00	0.80	0.50 (0.17)	0.83	0.95	0.74
GLasso	-	3.48 (1.30)	0.78	0.86	0.71	3.83 (0.77)	0.50	0.97	0.34
CLIME	-	1.88 (0.92)	0.79	1.00	0.65	2.98 (0.51)	0.51	1.00	0.34
ANT*	-	1.26 (0.98)	0.88	0.86	0.90	2.10 (0.72)	1.00	1.00	1.00
Hyperparam.		$(50, 1)$ – $(3, 1)$ – $(3, 1, 50, 1)$				$(100, 1)$ – $(20, 1)$ – $(20, 1, 50, 1)$			

TABLE 3.1 : Medians of the mean squared prediction errors (with standard deviations), F-scores, precisions and recalls after $N = 100$ repetitions of Scen. 0 to Scen. 6 ($N = 50$ for Scen. 4 and Scen. 6), with $n_e = 400$ and uncorrelated predictors. The suffix -or is used to denote 'oracle' settings. The hyperparameters chosen for the prior spike probability are indicated in the last row of each table, from left to right: (a, b) for (s) and (gs), (a_1, b_1, a_2, b_2) for (sgs).

due to the fact that it does not take into account the group structure, but still better than the penalized methods. However, the computational times involved (see remarks below) make (s) less relevant than (sgs) in these situations, even if the results are not drastically different. Unsurprisingly, (gs) is not suitable in the sparse-group-sparse settings in terms of support recovery. Our experiments show that it is able to identify influential groups without being mistaken but, even though the resulting estimates are small where they should be zero, it is not designed to be used for bi-level selection. Figure 3.1 shows that

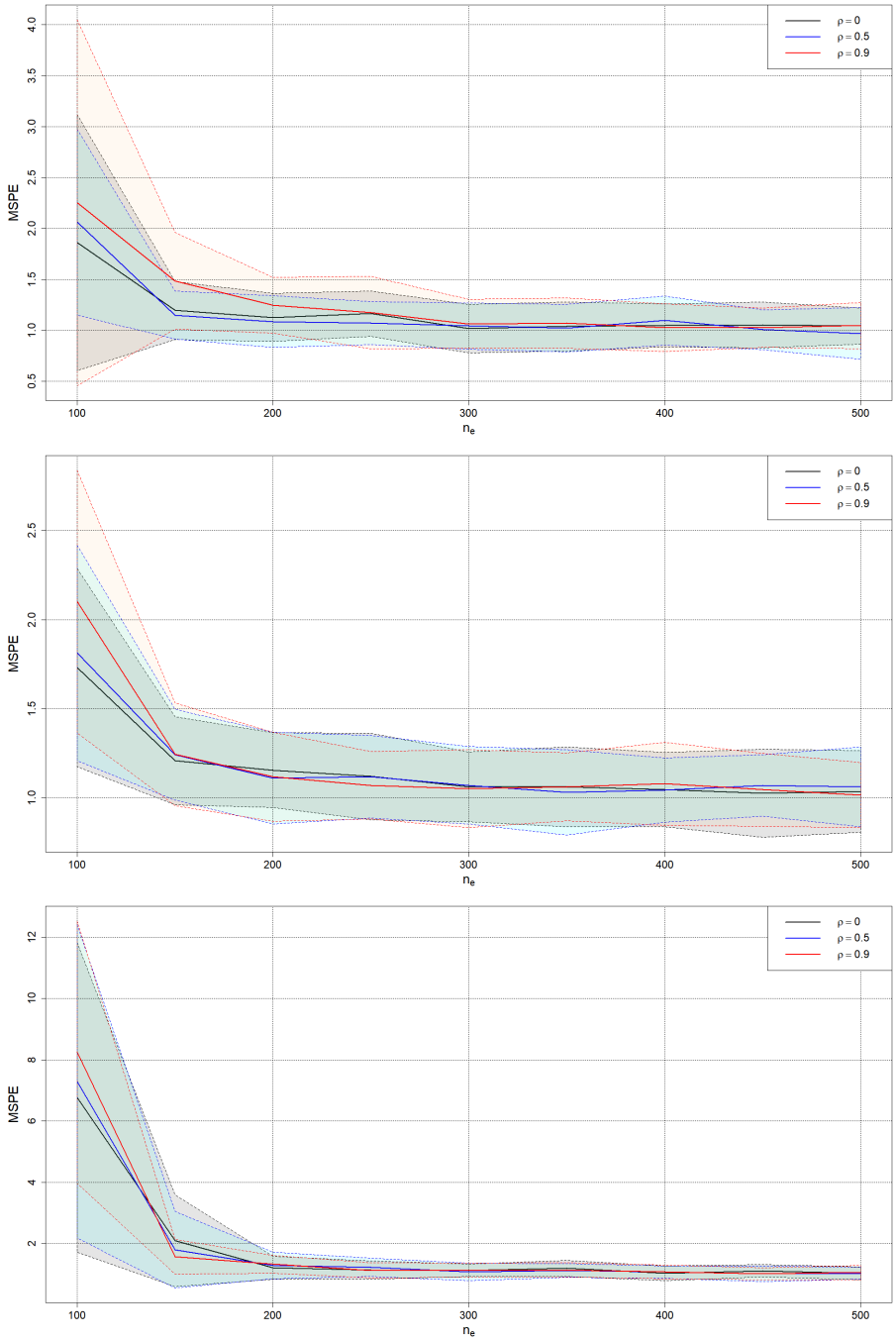


FIGURE 3.1 : Medians of the mean squared prediction errors obtained after $N = 100$ repetitions of Scen. 1 (top), Scen. 3 (middle) and Scen. 5 (bottom) with ± 1 standard deviation and n_e growing from 100 to 500. The black curves correspond to uncorrelated predictors ($\rho = 0$) while the blue and red curves correspond to correlated predictors ($\rho = 0.5$ and $\rho = 0.9$, respectively).

the results are pretty stable from $n_e = 200$ observations in the learning set: for $n_e < 200$ the MSPEs are rather chaotic before stabilizing. The same figure also highlights that the presence of correlation in

the predictors do not seem to have a significant effect on the estimation procedure, except for small size samples and high correlation where the degradation is noticeable. Overall, the real strength of the Bayesian spike-and-slab approach is clearly the support recovery of the direct links between predictors and responses but it seems that one can hardly expect to deal with very high-dimensional studies as long as we do not impose a group structure or a huge degree of sparsity. The highly competitive MSPEs obtained confirm the relevance of Bayesian PGGMs not only for variable selection but also for prediction purposes in the context of high-dimensional regressions.

3.6.2 Identification of a sparse set of predictors in a real dataset

Let us now study the `Hopx` dataset, fully described in [PBL+10]. It contains $p = 770$ genetic markers spread over $m = 20$ chromosomes from $n = 29$ inbred rats. It also contains the corresponding measured gene expression levels of $q = 4$ tissues (adrenal gland, fat, heart and kidney). The goal is to identify a sparse set of predictors that jointly explain the outcomes, with the natural group structure formed by chromosomes (see Table 3.2). This dataset has already been analyzed in [LBC+16], using a Bayesian regression without group structure, and later in [LMPS17] including group and sparse-group structures. So the PGGM is supposed to bring new perspectives about relationships in terms of partial correlations. A particularity of this dataset is that the responses are very correlated, so we should expect an estimation of Ω_y^{-1} with significant non-diagonal elements and a clear advantage in using PGGMs. Indeed, a predictor considered to be influencing all the outcomes could be the result of a direct relation to one tissue propagated to the others by an artificial correlation effect. As can be seen on Figure 3.2, the predictors are also highly correlated with their neighbors (for the sake of readability, we only represent the correlogram of predictors located on chromosomes 8, 9 and 10).

Chr.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Nb.	74	67	63	60	39	45	52	43	31	51	21	26	33	22	15	27	18	30	34	19

TABLE 3.2 : Number of markers on each chromosome, which correspond to the sizes κ_g of each group for $1 \leqslant g \leqslant 20$ when running (gs) and (sgs).

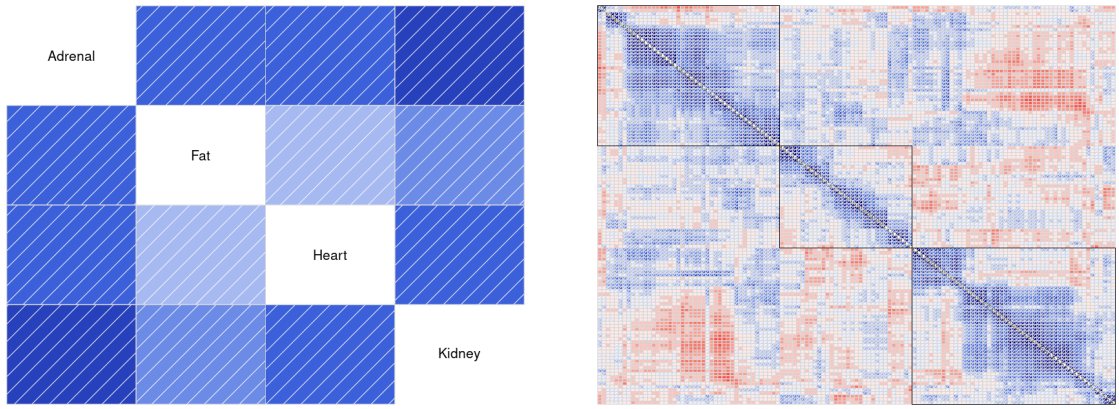


FIGURE 3.2 : Correlogram of responses (left) and correlogram of predictors located on chromosomes 8, 9 and 10 (right). The colormap associates red with negative correlations and blue with positive correlations.

The small sample size relative to the number of covariates (29/770) weakens the study. To strengthen our conclusions, we decided to run $N = 100$ experiments based on 25 randomly chosen observations and to aggregate the results. We first investigate the selection of predictors at the chromosomes scale, *i.e.* we run (gs) according to the previous protocol with an adaptative shrinkage and we choose $(a, b) = (1, 20)$ in the prior probability π. The empirical distribution of the posterior probability of inclusion for each chromosome is represented on the left of Figure 3.3. The selection procedure focuses on chromosomes 14, 15 and 17 (and not just on chromosomes 2 and 3 as in [LMPS17]) but the estimation process gives an overwhelming advantage to chromosome 14, far ahead of its neighbors. This is undoubtedly the influence of `D14Mit3`, a marker located on chromosome 14 and known to have a very significant effect on this dataset. The main conclusion to be drawn at this stage is that chromosome 14 has a positive effect on `Fat` and a *negative* effect on `Heart`, as can also be seen on the right of Figure 3.3. Therefore, it is likely that the overall positive influence of `D14Mit3` identified by previous authors is due to the combination of a direct positive link with `Fat`, a direct negative link with `Heart` and a correlation effect from the outcomes. This hypothesis is given additional credibility by the numerical results: from (gs), the corresponding column of Δ is approximately $(0.00, 0.04, -0.09, 0.00)$ which, through relations (1), leads to $(0.15, 0.25, 0.34, 0.21)$ as estimated regression coefficients. This roughly corresponds to the values indicated in Table 2 of [LMPS17], at least for the main effect on `Heart`. Thus for chromosome 14, the numerical results coincide but the interpretations are clearly different. Of course, similar reasonings can be carried out for the less influent chromosomes.

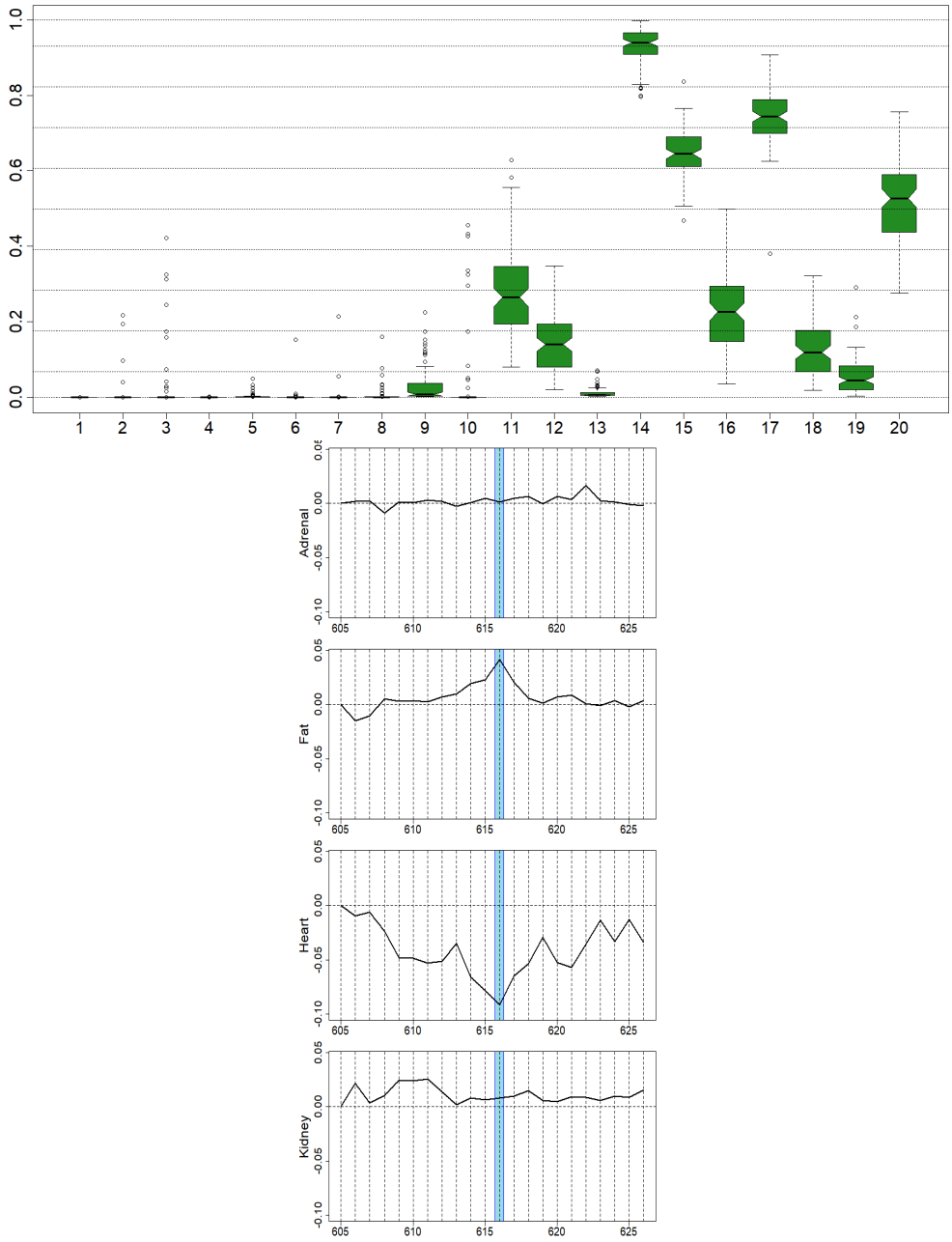


FIGURE 3.3 : Empirical distribution of the posterior probability of inclusion estimated by (gs) for each chromosome (left). Aggregated (gs) estimation of Δ on chromosome 14 with `D14Mit3` hilighted (right).

It is perhaps more interesting to look for a bi-level selection in order to identify a sparse set of markers and not only chromosomes. In this regard, (sgs) is launched using the same statistical protocol, adaptative shrinkage and hardly informative hypermarameters $a_1 = 3$, $b_1 = 1$, $a_2 = 1$ and $b_2 = 1$ which happen to be sufficient to generate a huge degree of sparsity. While many chromosomes are excluded from the model given by (gs), with (sgs) we see some contributions localized in certain chromosomes having little

influence when taken as a whole. At the markers scale, the randomness of the sampler and the high level of correlation between close predictors probably explain the presence of artifacts which sometimes make it difficult to distinguish the real contributions from the background noise. We therefore use the $N = 100$ experiments to build 95% confidence intervals and keep only significant estimates. By way of example, Figure 3.4 displays the results obtained on chromosomes 7, 8 and 14. The main markers standing out are summarized in Table 3.3 together with the kind of direct influences detected. Markers already highlighted in [LBC+16] or [LMPS17] are also indicated. One can see that most of our conclusions coincide, but new markers are suggested (especially on chromosome 8) and others have disappeared. Overall, the more stringent statistical protocol that we used led to the retention of fewer predictors with more guarantee. An important consequence of this study is the new interpretations in terms of direct influences allowed by PGGMs. Especially as the residual correlations, hidden in the estimation of $R = \Omega_y^{-1}$ and closely related to the correlations between the responses, are very high (greater than 0.7), as we suspected from Figure 3.2.

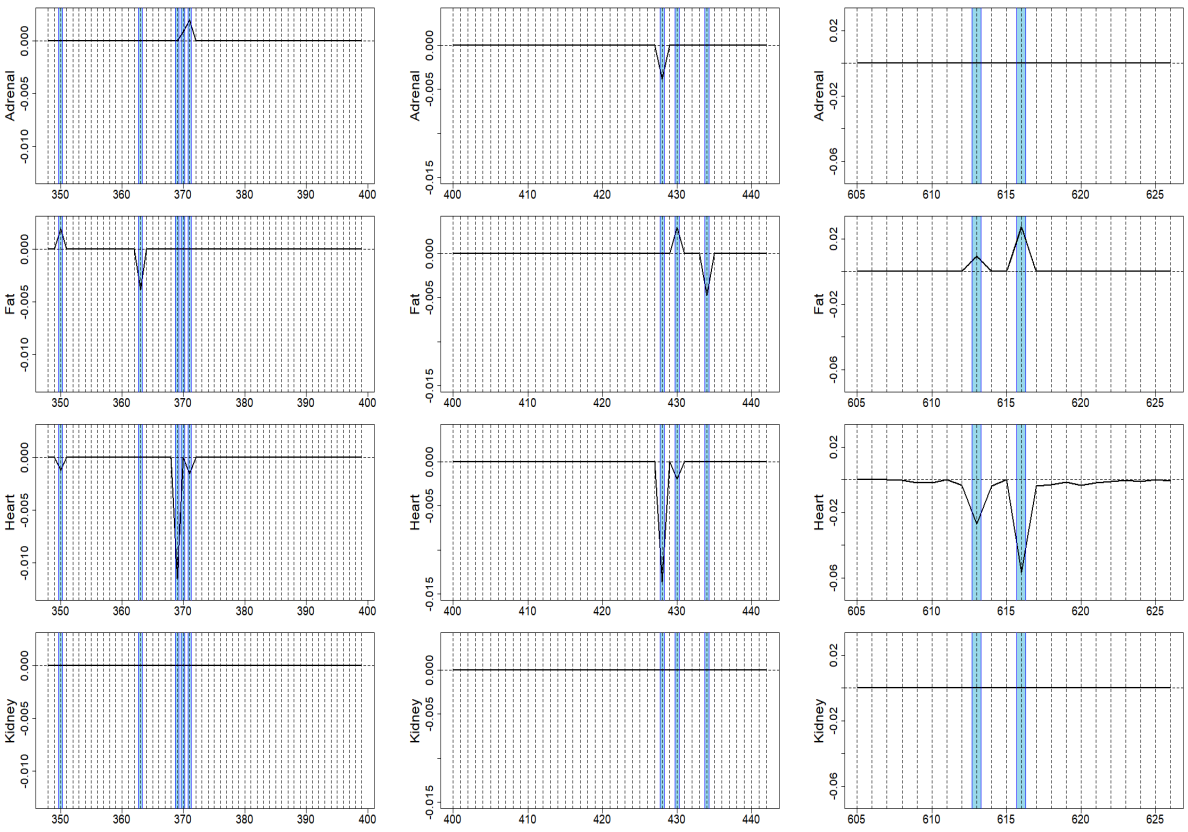


FIGURE 3.4 : Aggregated (sgs) estimation of Δ on chromosomes 7, 8 and 14, from left to right. The hilighted markers are `D7Cebr205s3`, `D7Mit6`, `D7Rat19`, `Myc` and `D7Rat17` for chromosome 7, `D8Mgh4`, `D8Rat135` and `Rbp2` for chromosome 8 and `D14Rat8` and `D14Mit3` for chromosome 14.

Chromosomes	Markers	Main direct influences
3	D3Mit16*	Adrenal+ Heart–
7	D7Cebr205s3*	Fat+ Heart–
	D7Mit6*	Fat–
	D7Rat19*	Heart–
	Myc*	Adrenal+
	D7Rat17	Adrenal+ Heart–
8	D8Mgh4	Adrenal– Heart–
	D8Rat135	Fat+ Heart–
	Rbp2	Fat–
10	D10Rat33*	Adrenal+
	D10Mit3*	Adrenal+
	D10Rat31*	Fat–
11	D11Rat47	Fat–
14	D14Rat8*	Fat+ Heart–
	D14Mit3*	Fat+ Heart–
15	D15Cebr7s13	Kidney–
	D15Rat21*	Adrenal+ Kidney–
17	Prl	Adrenal– Kidney–
20	D20Rat55	Kidney–

TABLE 3.3 : Main relations detected by (sgs). X* means that marker X has already been suggested by previous authors in this dataset. Y– (Y+) means that response Y is negatively (positively) influenced by X.

3.6.3 Discussion and Conclusion

To conclude, we would like to draw the attention of the reader to some weaknesses of the study, still under investigation. On the one hand, as soon as p is large (say, $p \geqslant 500$), the Bayesian studies should be conducted with a group structure or by promoting very sparse settings because due to the outline of the sampler, looping over each column of Δ may quickly become intractable. A group structure limits the number of loops (only $m \ll p$ per sampler iteration), although each loop may require the generation of large Gaussian vectors (up to $(q \times \kappa_g)$-dimensional), so compromises are needed. Subdividing the dataset is natural when it is intrinsically equipped with a group structure (*e.g.* that of the previous section), we could suggest otherwise a clustering of the set of predictors to gather similar entries and control the size of the groups. At this stage, our procedures cannot compete with the Lasso-type algorithms (GLasso, CLIME or even ANT) in terms of computational times. This is an issue on which future studies should focus (ongoing works are devoted to translating the samplers into more efficient environments), enhan-

ced MCMC methods may also be useful or novel computational strategies like the 'shotgun' stochastic algorithm of [YN20]. On the other hand, the procedures are obviously very sensitive to the initialization of the sampler, especially when $p \gg n$. For example, the term $|I_{\kappa_g} + \lambda_g \underset{\sim}{X}_g^t \underset{\sim}{X}_g|$ is likely to explode when κ_g is large and $\lambda_g > 1$, that is why λ_g has to be carefully controlled *via* an accurate initial choice of ℓ_g. Our heuristic approach is to initialize ℓ_g such that $\mathbb{E}[\lambda_g] < 1$ to control the behavior of $|I_{\kappa_g} + \lambda_g \underset{\sim}{X}_g^t \underset{\sim}{X}_g|$ during the first iterations. This works pretty well in practice, but needs to be done on a case-by-case basis, which could be improved. From a theoretical point of view, we should obviously enhance the estimation procedure by sampling from the $\mathcal{MGIG}_q$ distribution for $q > 1$, and not using the mode. Our fallback solution gives satisfactory but not completely rigorous results. In addition, it could be interesting to generalize the support recovery guarantee of Proposition 3.3.3 to (sgs), which is certainly possible at the cost of a few additional developments. Overall, our study shows that for the moderate values of p (up to 10^3 or 10^4), the Bayesian approach of the partial Gaussian graphical models is a very serious alternative to the frequentist penalized estimations, for prediction but also and especially for support recovery.

Chapter 4

Introduction to survival analysis

Survival analysis is a branch of Data Science that studies the time elapsed until the occurrence of a binary event, over a fixed observation period. Due to the convenience of this method to address an issue present in various research areas, it is used in several disciplines. In economics the event of interest could be the bursting of a speculative bubble, in industry the occurrence of a machine breakdown, and in sociology the obtaining of the first post-graduation job. Although for this same type of analysis the names differ according to the field (failure time analysis, reliability analysis, duration modeling, etc.), the questions treated and the statistical tools used are similar.
In the medical field, survival analysis exploits data from cohort surveys or clinical trials in order to extract information likely to help in designing effective therapeutic follow-up of patients. For example, it will be possible to study the impact of a treatment on the remission time of individuals, to extract the variables that may play an important role in the diagnosis, or to classify individuals according to their level of risk of recurrence [Chr87][ZCa16].

In this chapter, we provide an introduction to survival analysis in the context of modeling Triple Negative Breast Cancer (TNBC). The notions presented here will serve as a support for understanding Chapter 5, which deals with the issue of feature selection in survival analysis. We will rely mainly on the books by Kleinbaum and Klein [KK10], Lee and Wang [LW14], and Quigley [O'Q08], as well as various articles that we will present in due course.

4.1 Background and terminology

We are in a medical context where the data processed come from cohort surveys. A cohort survey can be divided into two phases. A first phase, known as the inclusion phase, during which patients who will serve as the basis of the study are recruited ; and a second phase, known as the follow-up phase, consisting of the therapeutic follow-up of patients and the collection of survey data until the event of interest occurs or the project ends. This type of survey produces very atypical data, in terms of survival time. First of all, the follow-up start date differs from patient to patient since it depends on the beginning of their exposure to risk (e.g., date of screening, start or end of treatment, etc.). In addition, since the disease progresses differently in each individual, the follow-up end date will also be specific to the patient. As shown in Figure 4.1, we can therefore observe three different situations.

1. The patient experienced the event during the project. This is the case for patients 3 and 5, their survival times are then observed.

2. The patient did not experience the event until the study ends. This is the case for patients 1 and 4, their survival times are not observed.

3. The patient's follow-up was interrupted during the study without them experiencing the event. This is the case of patient 2, they represent individuals who leave the project for personal reasons (*e.g.* moving, voluntary discontinuation of treatment, death not related to the disease), they are considered **lost to follow-up** and their survival times are not observed.

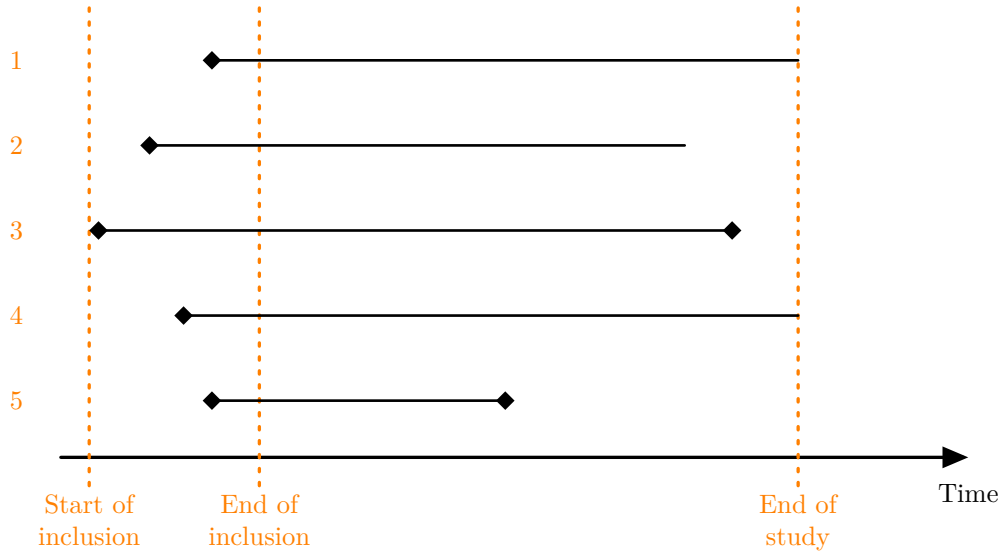


FIGURE 4.1 : Illustration of survival data in a study-time scale. Observed survival times are indicated by solid diamonds, and the others are censored observations.

In this chapter, we present tools that can exploit these heterogeneous data to derive information that may help in therapeutic decision-making. Before going any further, let us review and define the vocabulary specific to cohort studies that we will be using in this introduction.

– **Event or death**: the terms death and event are synonymous in survival analysis and simply refer to the phenomenon of interest.

– **Start of inclusion**: refers to the beginning of the study project ; the date from which the patients are recruited.

– **End of inclusion**: refers to the date on which recruitment of new patients ends. The individuals who joined the project during the inclusion period constitute the final cohort sample.

– **End of study**: refers to the closing date of the trial, and therefore the end of patient follow-up. Any event occurring after this date will not be observed.

– **Entry date**: corresponds to the follow-up start date of an individual, in other words their T_0. This date must correspond to the beginning of the individual's exposure, it is important to determine

the event with which it is associated in order to have a good interpretation of the study results. In the case of breast cancer, this may be the date of cancer screening, a medical examination, the start of treatment or a surgical operation.

- **Date of event**: corresponds to the date on which the event occurred.
- **Date of last follow-up**: corresponds to the date of the last news of an individual lost to follow-up, *i.e.* who left the clinical trial before the event occurs.
- **Survival time**: corresponds to the time elapsed between the entry date of an individual and their date of event.

4.2 Survival model

The main idea of survival analysis is to determine the impact of individual characteristics on survival time. In the context of breast cancer, this would allow us to determine, for example, which patient profiles have a high probability of remission, and conversely which patient profiles will be more exposed to recurrence or death, and why. We might think of a standard linear regression problem defined as follows

$$T_i = \beta^t X_i + \epsilon_i,$$

where T_i corresponds to the survival time of individual i, $X_i \in \mathbb{R}^p$ the values by individual i on the p predictor variables, $\beta \in \mathbb{R}^p$ the vector of regression coefficients, and ϵ_i a noise term. However, we have seen previously that cohort survey data may include individuals who did not experience the event during their follow-up. Since their survival time is not observed, this creates a problem of missing informative data on the variable T. In this case, the use of standard linear regression is therefore excluded. An intuitive idea to get around this problem would be to create the regression model using only the individuals whose survival time is reported. This idea cannot be retained because it would introduce a significant bias in the study results. First, it means working under the assumption that the event will be observed for all individuals in the population, but it also implies a loss of information about the omitted individuals. Indeed, although the latter did not observe the event during the survey, we know that they survived at least up to some observed time. This is important information that survival analysis models can exploit by introducing the notion of censoring.

4.2.1 Censoring

Censoring occurs when we have some information on the survival of an individual, but the date of death is not explicitly known. The idea is to observe not only the realizations of the survival time variable T, but also the quadruplet (Y, C, T, δ), where for each individual i, we will denote Y_i their observed survival time (follow-up time), C_i their censoring time, T_i their real survival time, and δ_i their event status. The definition of the variables Y and C depends on the type of censoring used. We present the three most widespread, their formulation is conditioned by the information previously available on individuals' survival time.

Left censoring

Left censoring applies when the patient's actual survival time is less than their observed survival time. This phenomenon occurs in particular when the start date of exposure is not observable, and in studies where patients may observe the event outside the observation period. An example that illustrates this first case is a screening test. The patient contracts the disease before the test results are available.

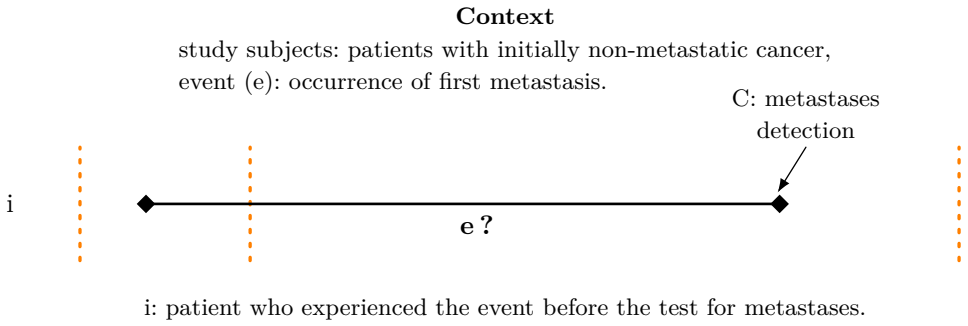


FIGURE 4.2 : Example of left censoring.

The variables Y and δ are then defined as follows:

$$Y = \max(T, C) \quad ; \quad \delta = \mathbb{1}_{\{T \geqslant C\}}.$$

Right censoring

Right censoring applies when the patient did not observe the event during their follow-up period. We then know that their actual survival time is greater than their observed survival time. This form of censoring is very common in analyses with cohort survey data, because even when the protocol provides for continuous monitoring, this is limited in time and some individuals leave the project before the event occurs. There are three subtypes of right censoring, in this report we will only discuss type 3 which is very common in clinical research.

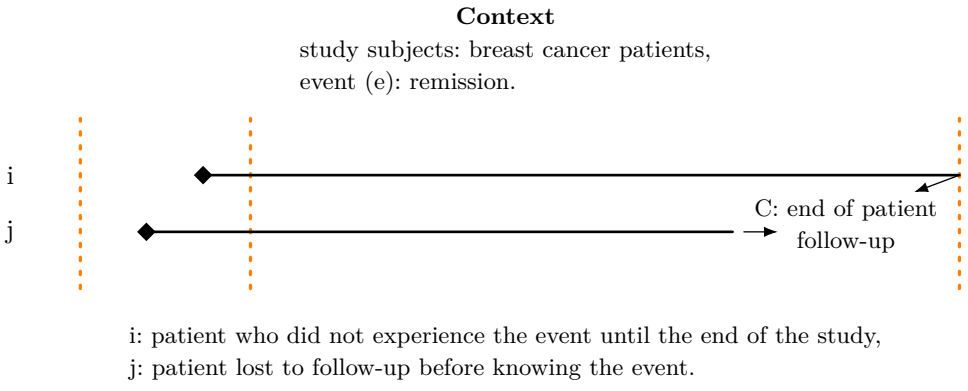


FIGURE 4.3 : Example of right censoring.

The variables Y and δ are then defined as follows:

$$Y = \min(T, C) \quad ; \quad \delta = \mathbb{1}_{\{T \geqslant C\}}.$$

Interval censoring

Interval censoring applies when the date of the event falls between two known dates. Like left censoring, it is found in studies where the monitoring is discontinuous ; patients can then observe the event between two observation periods. Interval censoring can be seen as a generalization of left and right censoring. It

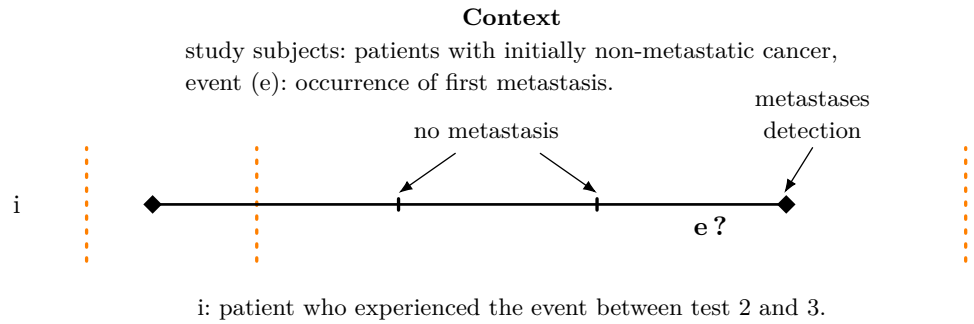


FIGURE 4.4 : Example of interval censoring.

also occurs in many clinical studies, but is ignored for convenience, assuming that the date of observation coincides with the date of the event [Ra18]. Otherwise Y is defined as follows:

$$Y = \begin{cases}]C_l; C_r] & \text{if } \delta = 1, \\ ]C_l; \infty[& \text{if } \delta = 0. \end{cases}$$

4.2.2 Survival time distribution

In survival analysis, estimating the probability distribution of the variable T is a crucial point. This one is characterized and can be defined by different functions (*i.e* survival function, density function, cumulative distribution function...), but we will see later on that the survival and hazard functions are the most manipulated in practice. From now on, we will assume that T is a continuous non-negative variable, and denote $f(t)$ its density function and $F(t)$ its cumulative distribution function.

The survival function

The survival function $S(t)$ models the probability of an individual to survive beyond a fixed time t. In other words, it returns the probability that the patient has not yet observed the event by this instant t. The survival function is defined by

$$S(t) = \mathbb{P}(T > t), \qquad \text{for } t > 0. \tag{1}$$

$S(t)$ is in theory a continuous and decreasing function decreasing from 1 to 0 when t goes from 0 to $+\infty$. However, in practice its estimate is discretized according to the training data, resulting in a step function. Figure 4.5 illustrates graphical representations of a function $S(t)$ and the empirical survival function $\hat{S}(t)$, commonly called survival curves.

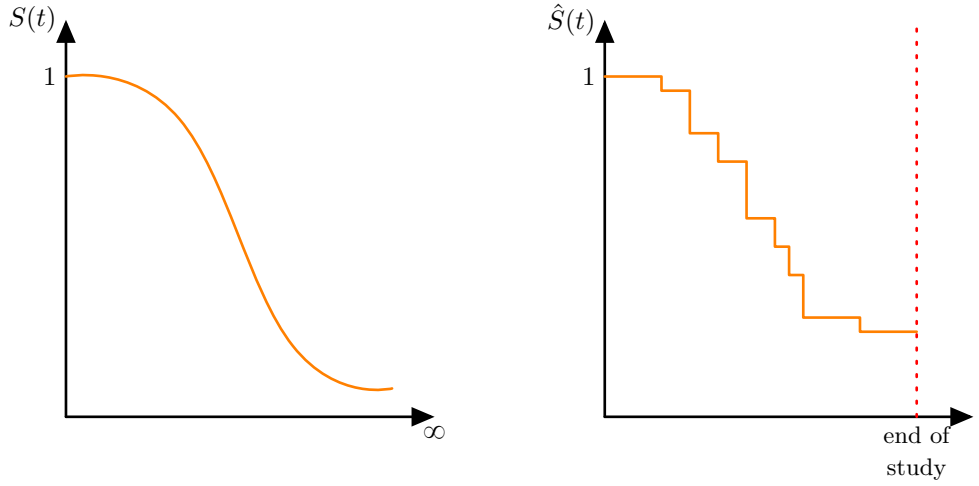


FIGURE 4.5 : Example of survival curves: (left) the theoretical curve - (right) the estimated curve.

The hazard function

The hazard function $h(t)$, also known as instantaneous risk rate, returns an indication of the probability of experiencing the event within a short time interval after t, given that the patient has survived up until t. In this sense, it makes it possible to determine the periods during which individuals are most exposed to the instantaneous risk of death, hence its name. The hazard function is defined as follows:

$$h(t) = \lim_{h \to 0} \frac{\mathbb{P}(t < T \leqslant t + h | T > t)}{h} = \frac{f(t)}{S(t)} = -\ln(S(t))', \qquad \text{for } t > 0. \tag{2}$$

Like with the survival curves, it is recommended to illustrate individuals' hazard functions graphically, to get a better interpretation of their instantaneous risk evolution over time. Figure 4.6 illustrates examples of hazard curves. We will see later that the pattern of these curves can give us some indications on the model distribution family, and thus conversely, it will be possible to make assumptions on the model law from prior information about the evolution of the risk over time.

The cumulative hazard function

The cumulative hazard function represents, by definition, the quantity of risk accumulated over a given period of time. It is written as

$$H(t) = \int_0^t h(u)du = -\ln(S(t)), \qquad \text{for } t > 0. \tag{3}$$

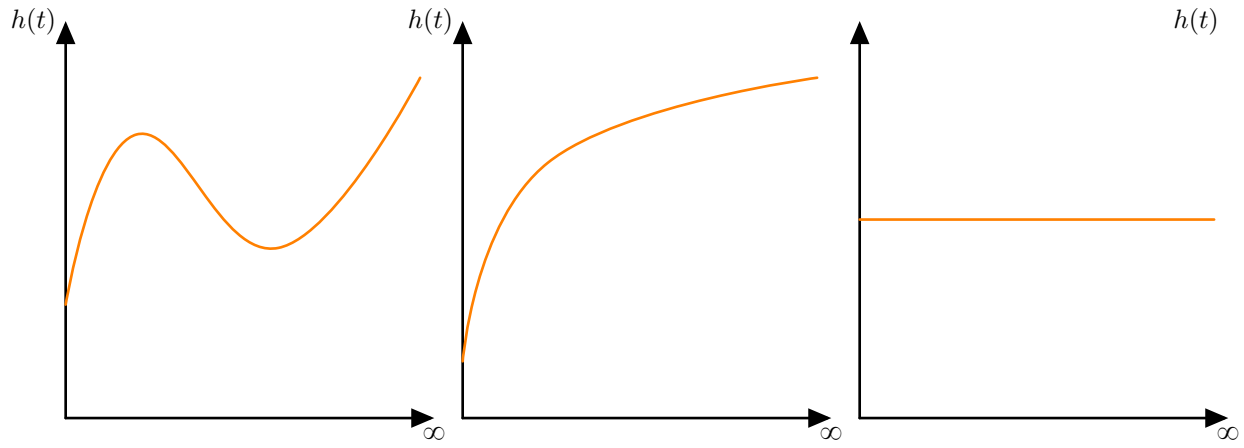


FIGURE 4.6 : Example of hazard curves.

We will now present some basic estimation methods, which are essential for this introductory chapter. For this purpose, let us consider now and in all the chapter, the sample of n independent observations $(\mathbf{Y}, \delta, \mathbf{X})$, where for each patient i we have the triplet (Y_i, δ_i, X_i). $Y_i \in \mathbb{R}^+$ is the observed survival time (*i.e.* real or censoring time), $\delta_i \in \{0, 1\}$ is the event status, and $X_i \in \mathbb{R}^p$ stands for the individual characteristics on the p predictor variables.

4.3 Non-parametric estimation - the Kaplan-Meier estimator

The Kaplan-Meier model, also called product limit estimator, provides a non-parametric estimate of the survival function $S(t)$. It is based on the fact that a patient will have a survival time greater than a fixed time t, if they were always at risk before t and did not experience the event at t. Thus, considering two times t_i and t_{i-1} such that $t_i > t_{i-1}$ and $t_{i-1} \geqslant 0$, we can see that

$$\begin{aligned}\mathbb{P}(T > t_i) &= \mathbb{P}(T > t_i, T > t_{i-1}) \\ &= \mathbb{P}(T > t_i | T > t_{i-1}) \mathbb{P}(T > t_{i-1}).\end{aligned}$$

Suppose we have a set of distinct ordered times of death (t_i), such that for $i = 1, ..., n$, $t_{i-1} < t_i$ and $t_0 = 0$, by induction we have

$$\mathbb{P}(T > t_k) = \prod_{i=1}^{k} \mathbb{P}(T > t_i | T > t_{i-1}), \qquad \text{with } k \in [\![1, n]\!]. \tag{4}$$

The Kaplan-Meier approach estimates the conditional probability of surviving at time Y_i by the proportion of patients still at risk at Y_i and who did not experience the event at that time. Let n_i be the number of individuals at risk at time Y_i (*i.e.* the number of patients present in the study at Y_i), and e_i the number of events at Y_i, we have

$$\hat{\mathbb{P}}_{KM}(T > Y_i | T > Y_{i-1}) = 1 - \frac{e_i}{n_i}.$$

Thus, under the assumption of distinct times Y_i, the Kaplan-Meier estimator is given by

$$\hat{S}_{KM}(t) = \prod_{i:Y_i \geqslant t} \left(1 - \frac{\delta_i}{n_i}\right) = \prod_{i:Y_i \geqslant t} \left(\frac{n-i}{n-i+1}\right)^{\delta_i}. \tag{5}$$

Under the assumption of independence between survival and censoring times, it is possible to show that $\hat{S}_{KM}(t)$ is a consistent maximum likelihood estimator with negligible bias [KM58].

The Kaplan-Meier estimator does not incorporate individual characteristics in its estimation procedure, which limits its scope of application. In practice, this method is used more for descriptive statistics purposes than for fitting a predictive model. In particular, Kaplan-Meier curves can be used to obtain a general idea of the evolution of the survival probability, or to compare survival among independent groups distinguished by a specific characteristic (*e.g.* the treatment received).

4.4 Semi-parametric estimation - Cox model

Semi-parametric estimation aims to build a reliable model for which a minimum of assumptions are made. In survival analysis and especially in medical field, the Cox proportional hazard model [Cox75] is by far the most widely used. Its popularity stems from its robustness which allows it to approximate the results of a properly constructed parametric model. We will see, however, that this flexibility requires compromises.

4.4.1 Proportional hazard models

Proportional hazard models express a multiplicative effect of predictors on the hazard function. The latter is defined as the product of a baseline hazard function that depends only on time, and a positive function which represents the impact of individual characteristics. It is given by

$$h(t|X) = h_0(t)\, R(\beta, X), \qquad \forall\, t > 0, \tag{6}$$

where β is the risk parameter to be estimated. Note that it is important to ensure the positivity of the function $R(\beta, X)$ over the range of possible values for β and X, in order to maintain the interpretability of the hazard function discussed in section 4.4.3. The hazard function thus defined presents a strong assumption that should be evaluated, namely the effect of explanatory variables does not vary over time. Thus, for any pair of individuals i and j, who take the values X_i and X_j on the predictors, the ratio of their hazard functions is constant over time. Equivalently, we may say that the risk associated with individual i is proportional to the risk associated with individual j, and the proportionality constant does not depend on time.

$$HR_{ij} = \frac{h(t|X_i)}{h(t|X_j)} = \frac{R(\beta, X_i)}{R(\beta, X_j)} \qquad \Longleftrightarrow \qquad h(t|X_i) = \frac{R(\beta, X_i)}{R(\beta, X_j)} h(t|X_j), \qquad \forall\, t > 0, \forall\, i, j.$$

These models take their name from this property. Furthermore, the hazard ratio is a measure of the predictor effects. It has a similar interpretation to the odds ratio in logistic regression. Let be a patient

i, a reference patient j, and their estimated hazard ratio $\widehat{HR}_{ij}$,

- if $\widehat{HR}_{ij} = 1$, there were no significant differences between the two patients,
- si $\widehat{HR}_{ij} > 1$, the risk for patient i is greater than for patient j,
- si $\widehat{HR}_{ij} < 1$, the risk for patient i is lower than for patient j.

Similarly, it enables to compare the risk between different groups of patients.

4.4.2 The Cox proportional hazards model

The Cox model is defined as follows

$$h(t|X) = h_0(t)\exp(\beta^t X), \qquad \forall\, t > 0. \tag{7}$$

Here the relative risk function $R(\beta, X)$ is written as the exponential of a linear expression between predictors and regression coefficients, and the baseline hazard function describes the risk for individuals with $X_0 = 0 \in \mathbb{R}^p$, who can serve as a reference. The hazard ratio for two individuals i and j is given by

$$\frac{h(t|X_i)}{h(t|X_j)} = \exp(\beta^t(X_i - X_j)), \qquad \forall t > 0. \tag{8}$$

Although it assumes a parametric form for the relative risk, the Cox model imposes no assumptions on the baseline, and focus on inference that allowed $h_0(t)$ to remain arbitrary. Therefore, its estimation is based on a semi-parametric procedure.

4.4.3 The partial likelihood

The likelihood function for right-censored data is built from two types of contribution. Suppose an individual i is observed for a time Y_i. If the individual i observed the event, *i.e* $Y_i = t_i$ and $\delta_i = 1$, their contribution to the likelihood is the density at Y_i, given by $f(Y_i|X_i, \beta) = h(Y_i|X_i, \beta)S(Y_i|X_i, \beta)$. However, if the individual i is censored, *i.e* $Y_i = C_i$ and $\delta_i = 0$, all we know is that they have survived beyond the observed time ; their contribution is then the probability of this *event*, which is $S(Y_i|X_i, \beta)$. Thus, the likelihood associated with a survival model is given by

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} f(Y_i|X_i, \beta)^{\delta_i} S(Y_i|X_i, \beta)^{1-\delta_i} = \prod_{i=1}^{n} h(Y_i|X_i, \beta)^{\delta_i} S(Y_i|X_i, \beta), \tag{9}$$

since $f(Y_i|X_i, \beta) = h(Y_i|X_i, \beta)S(Y_i|X_i, \beta)$ (see (2)). To apply the standard maximum likelihood method, and estimate the regression coefficients β in the context of proportional hazard models, we would need prior knowledge on the baseline hazard function. To overcome this restriction, Cox proposed a partial likelihood, based on the conditional probabilities of death, which eliminates the term h_0. The idea is to recover the order of event occurrences rather than their distribution. Thus the partial likelihood will be the product on the observed data, that the individual i knows the event at time Y_i, among the individuals

still at risk at Y_i, and given that there was an event at Y_i. Which give

$$\mathcal{L}_p(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta X_i)}{\sum_{j \in R(Y_i)} \exp(\beta X_j)} \right)^{\delta_i} = \prod_{i:\delta_i=1} \frac{\exp(\beta X_i)}{\sum_{j \in R(Y_i)} \exp(\beta X_j)}, \tag{10}$$

where $R(Y_i) = \{j : Y_j \geqslant Y_i\}$ is the set of individuals still at risk at time Y_i. The partial likelihood only partially considers censored individuals. In this sense, it cannot be considered as a genuine likelihood, however it behaves as such and has good computational capacities [KK10]. The estimation of the coefficients $\hat{\beta}$ is generally obtained by maximizing the partial log-likelihood using a gradient descent algorithm.

4.5 Parametric estimation - some usual distributions

The Cox model semi-parametic estimation allows to recover the effects of predictor variables, without having to estimate the baseline function. Nevertheless, this flexibility comes with an inability for the method to recover the model survival distribution. Parametric models approximate the complete survival distribution while incorporating the effects of predictor variables in the fitted model. To do this, these methods rely on the strong assumption that the survival function follows a known distribution, the aim being to estimate the parameters of this distribution. Here we will present some well-known distribution families and formalize their estimation in the framework of proportional hazards models.

4.5.1 Some distribution families

The first step in the estimation procedure is based on the choice of the model distribution family. When the practitioner has no prior idea of the distribution of the data, the law of the model can be presumed from the form of one of the five functions which define the distribution of T, *i.e.* S, f, F, h and H. For example, it will be possible to choose a distribution family, according to the form of the survival function estimated by a non-parametric method, or even according to the assumptions made about the behavior of the risk over time. The Table 4.1 represents the density, survival and hazard functions of the most widespread distribution families in the literature, and the Figure 4.7 illustrates the hazard and survival curves associated with them according to the values taken by their parameters.

Distribution	**Density** $f(t)$	**Survival** $S(t)$	**Hazard** $h(t)$	**Parameters**
Exponential	$\lambda \mathrm{e}^{-\lambda t}$	$\mathrm{e}^{-\lambda t}$	λ	rate: $\lambda > 0$
Weibull	$k\lambda\,(t\lambda)^{k-1}\,\mathrm{e}^{-(t\lambda)^k}$	$\mathrm{e}^{-(t\lambda)^k}$	$k\lambda\,(t\lambda)^{k-1}$	shape: $k > 0$; scale: $\lambda^{-1} > 0$
Gamma	$\frac{\theta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} \mathrm{e}^{-\theta t}$	$1 - \frac{\gamma(\alpha,\theta t)}{\Gamma(\alpha)}$	$\frac{f(t)}{S(t)}$	shape: $\alpha > 0$; rate: $\theta > 0$
Log-logistic	$\frac{(\theta/\alpha)(t/\alpha)^{\theta-1}}{(1+(t/\alpha)^\theta)^2}$	$\frac{1}{1+(t/\alpha)^\theta}$	$\frac{(\theta/\alpha)(t/\alpha)^{\theta-1}}{1+(t/\alpha)^\theta}$	shape: $\theta > 0$; scale: $\alpha > 0$

TABLE 4.1 : Characteristic functions of the survival time for the usual distribution families.

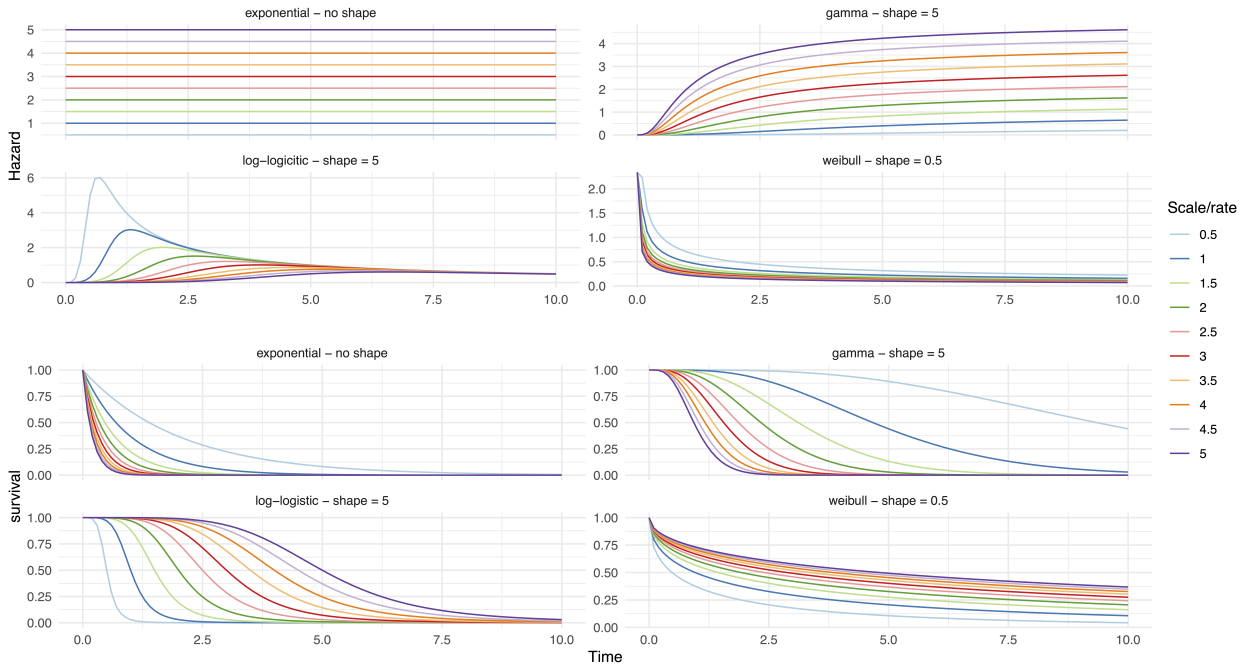


FIGURE 4.7 : Representation of the hazard and survival functions of the usual distribution families according to shape, rate and scale parameters.

We can see that an exponential-type model is characterized by a constant risk over time and a clear decrease in the survival function. These assumptions may hold for example when studying the lifetime of some kinds of electronic components, but are generally unrealistic in biomedical field. A Weibull model is, on the other hand, characterized by a monotonic failure rate and is therefore suitable for many problems.

In the absence of explanatory variables, parametric models can be estimated by maximum likelihood. Following the same reasoning as in Section 4.4.3, this likelihood is given by

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(Y_i)^{\delta_i} \, S(Y_i)^{1-\delta_i}, \tag{11}$$

where θ is the parameters characterising the distribution of T. The estimation $\hat{\theta}$ of the parameters θ is usually obtained using iterative optimisation procedures such as the Newton-Raphson algorithm.

4.5.2 Proportional hazard models

In order to incorporate individual characteristics into the estimation procedure, a natural approach is to express the parameters of the desired distribution in terms of the covariates. There is three popular methods to modelling survival data with covariates: the Proportional Hazard (PH) model, Accelerated Failure Time (AFT) model and the Proportional Odds (PO) model. For the sake of brevity, we will only present the parametric PH model. We direct the interested reader to the books by Kleinbaum and Klein [KK10], and Lee and Wang [LW14], which present the other two approaches.

The parametric version of PH models is obtained by assuming a parametric form on the baseline hazard $h_0(t)$ of the hazard function (7). Among the distribution families presented in Table 4.1, only the exponential and Weibull distributions can accommodate the proportional hazard assumption. Suppose that T follows a Weibull distribution of parameters (λ, k). When building a PH model, an immediate assumption for the baseline hazard would be

$$h_0(t) = \lambda k \left(\lambda t\right)^{k-1}. \tag{12}$$

Multiplying this term by the relative risk related to individual characteristics, we obtain the hazard function of the PH model

$$h(t) = \lambda k \left(\lambda t\right)^{k-1} \exp(\beta X). \tag{13}$$

By setting $\lambda^\star = \lambda \exp(\frac{\beta X}{k})$, one can show that we obtain again a Weibull distribution with parameters $(\lambda^\star, k)$. Thus, the Weibull family is closed under the proportional hazard assumption. Obviously, this reasoning holds for the exponential family, since it is a special case of Weibull where $k = 1$. The estimation of the parameters of these models is obtained by maximizing the likelihood using iterative optimisation procedures such as the Newton-Raphson method.

4.6 Estimation of penalized models

Although the Cox model is highly appreciated in the literature, it is not the best suited to high-dimensional problems, where the sample size is much smaller than the number of explanatory variables, and for which there is a non-negligible amount of censored observations. To avoid falling into overfitting, researchers have introduced penalized models. For most of the methods encountered in the literature, the principle is basically the same as the penalized methods presented earlier in this manuscript. Starting from the initial optimization problem, we incorporate a penalty term to the objective function, inducing sparsity and shrinkage in the coefficient vector. The survival Lasso of Tibshirani [Tib97] is again the best known method. It consists in penalizing the partial log likelihood of the Cox model with an ℓ_1-norm penalty. The new optimisation problem is given by

$$\hat{\beta} = \arg\min_{\beta} \big(- \ell\ell_p(\beta) + \lambda \left|\beta\right|_1 \big). \tag{14}$$

Most of the penalized estimation methods in survival analysis are based on the Cox model, and differ in the form of penalization proposed. The literature includes a wide variety of them adapting to different constraints of high dimensional survival analysis. In Chapter 5, we will present some of them.

4.7 Proportional hazards Cure model

The standard survival analysis is based on the strong assumption that all individuals will eventually observe the event of interest if the follow-up time is long enough. However, depending on the problem studied, it is likely that a part of the population is no longer at risk and will never experience the event. In the context of breast cancer, for example, taking an effective treatment can spare some patients

from the recurrence event. These phenomena are modeled using Cure models, which assume that a fraction of the censored observations are in fact "cured" to the event. In other words, we will assume that $\lim_{t\to+\infty} S(t) > 0$, which is characterized by a survival curve that reaches a plateau due to the absence of new events when there is only the cured population left to follow. Thus, a survival curve, estimated by a nonparametric method, that shows a long and stable plateau with a strong censoring at the tail (as in Figure 4.8), can be considered as an empirical evidence of cured individuals in the population. Many parametric, non-parametric and semi-parametric methods of Cure models have been proposed in the literature [Gol84, PD00, Far82]. We focus here on the Proportional Hazards Cure model.

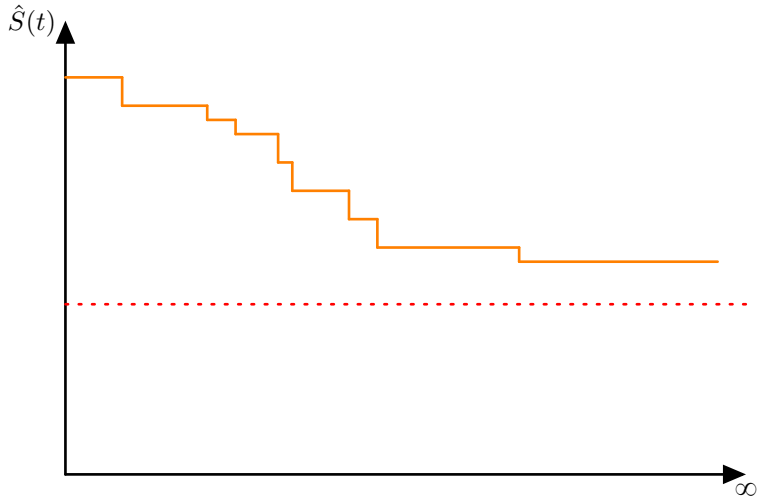


FIGURE 4.8 : Example of an estimated survival curve adapted to cure models.

4.7.1 Model formulation

An intuitive way to formulate a Cure model is to use a mixture modeling approach, where the study population is assumed to be a mixture of cured and exposed individuals. Let L be the dummy variable that indicates whether an individual will eventually experience the event ($L = 1$) or not ($L = 0$), and define $\pi(X) = \mathbb{P}(L = 1 \,|\, X)$ the fraction uncured of the population having characteristics X. Under the mixture modeling approach, the survival function now satisfies

$$S^{(c)}(t \,|\, X) = (1 - \pi(X)) + \pi(X)\, S(t \,|\, L = 1, X), \tag{15}$$

where $S(t \,|\, L = 1, X)$ denotes the survival function for the exposed group. Thus, to study the predictor effects on the cure rate and survival distribution, we only need to model the distribution of L and the conditional distribution $T|L = 1$. In this vein Farewell proposed a parametric approach [Far82]. For the incidence part, they applied a logistic form to the distribution of L, *i.e.*

$$\pi(X) = \mathbb{P}(L = 1 \,|\, X, b) = \frac{\exp(b^t X)}{1 + \exp(b^t X)}\,, \tag{16}$$

where $b \in \mathbb{R}^p$ is the vector of coefficients for this part. Then for the latency, they assumed the survival time for the exposed fraction to have a Weibull distribution. Kuk and Chen subsequently proposed a generalization of the Farewell model, which consists in estimating the conditionnal distribution $T \,|L = 1$ using a Cox model [KC92]. The associated conditional hazard and survival functions are given by

$$h(t \,|\, L = 1, X) = h_0(t \,|\, L = 1) \, \exp(\beta X), \tag{17}$$

$$S(t \,|\, L = 1, X) = S_0(t \,|\, L = 1)^{\exp(\beta X)}, \tag{18}$$

where $S_0(t \,|\, L = 1) = \mathbb{P}(T > t \,|\, L = 1, X = 0)$ is the baseline conditional survival function and $\beta \in \mathbb{R}^p$ the vector of coefficients for the latency part.

4.7.2 Maximum likelihood estimation

Building a PH mixture Cure model relies on the estimation of the parameters b and β, which is usually done by likelihood maximization. Suppose an individual i is observed for a time Y_i. If the individual i observed the event, *i.e* $Y_i = t_i$ and $\delta_i = 1$, we know they are not in the cured fraction and therefore $L_i = 1$. Their contribution to the likelihood is given by $\pi(X_i)\, f(Y_i \,|\, L_i = 1, X_i)$. However, if the individual i is censored, *i.e* $Y_i = C_i$ and $\delta_i = 0$, we cannot tell if they are exposed and will experience the event at some future time after their follow-up, or if they are cured. Their contribution is the expression of the cure survival function at Y_i, which is $\pi(X_i) + (1 - \pi(X_i))\, S(Y_i \,|\, L_i = 1, X_i)$. Thus, the observed full likelihood is given by

$$\begin{aligned}
\mathcal{L}(b, \beta) &= \prod_{i=1}^{n} \Big[\pi(X_i)\, f(Y_i \,|\, L_i = 1, X_i, \beta)\Big]^{\delta_i} \Big[\pi(X_i) + (1 - \pi(X_i))\, S(Y_i \,|\, L_i = 1, X_i, \beta)\Big]^{1-\delta_i}, \\
&= \prod_{i=1}^{n} \Big[\pi(X_i)\, h(Y_i \,|\, L_i = 1, X_i, \beta)\, S(Y_i \,|\, L_i = 1, X_i, \beta)\Big]^{\delta_i} \Big[\pi(X_i) + (1 - \pi(X_i))\, S(Y_i \,|\, L_i = 1, X_i, \beta)\Big]^{1-\delta_i}.
\end{aligned} \tag{19}$$

When the cured fraction satisfies $1 - \pi(X_i) = 0$ for all combinations of characteristics X_i, we recover the Cox model and the coefficients β can be estimated by maximizing the associated partial log-likelihood. However, in the presence of a cured fraction, we cannot isolate the conditional survival function as performed in the Cox likelihood. In any case, it is not recommended to remove $S_0(t \,|\, L_i = 1)$ from the optimization problem, at the risk of losing information about b [ST00]. Many estimation methods have been proposed to address this problem [ST00, PD00, KC92, Lu08]. We can cite as an example the non-parametric estimation approach of Peng and Dear [PD00] and Sy and Taylor [ST00]. The idea is to express the complete likelihood with the latent variable L. Since L_i is unobserved when $\delta_i = 0$, there are two possibilities. Either the subject is in the cured fraction and contribution to the likelihood is the probability to be cured, either they are exposed and their contribution is given by the conditional survival

function. Thus, the complete and unobserved likelihood is given by

$$\begin{aligned}
\mathcal{L}(b,\beta) = & \prod_{i=1}^{n} \Big[\pi(X_i)\, h(Y_i|L_i=1,X_i,\beta) S(Y_i|L_i=1,X_i,\beta)\Big]^{\delta_i L_i} \\
& \times \prod_{i=1}^{n} \Big[\pi(X_i)\, S(Y_i|L_i=1,X_i,\beta)\Big]^{(1-\delta_i)L_i} \\
& \times \prod_{i=1}^{n} \Big[1-\pi(X_i)\Big]^{(1-\delta_i)(1-L_i)}, \qquad (20) \\
= & \prod_{i=1}^{n} \pi(X_i)^{L_i}\ (1-\pi(X_i))^{(1-L_i)} \\
& \times \prod_{i=1}^{n} \Big[h(Y_i|L_i=1,X_i,\beta) S(Y_i|L_i=1,X_i,\beta)\Big]^{\delta_i L_i} \Big[S(Y_i|L_i=1,X_i,\beta)\Big]^{(1-\delta_i)L_i}. \qquad (21)
\end{aligned}$$

The complete-data full likelihood is expressed as the product of two elements, each including one of the two parameters of the mixture model. The estimation procedure is performed with an EM algorithm, first by computing the expected complete-data likelihood with respect to L, then by maximazing the likelihood obtained.

Chapter 5

Stepwise Variable Selection for Survival Analysis

Multi-omics analysis involves very high-dimensional data, where the number of individuals is considerably smaller than the number of explanatory variables. In particular, depending on the omics measures considered, the number of predictors can reach an order of magnitude of 10^5, whereas cohort surveys generally contain only a few hundred individuals. This second point can be explained by the various constraints related to the recruitment of individuals within the surveys, but also by the cost and time constraints required for the collection of information, which are proportional to the number of omics approaches considered and to the quantity of individuals selected. In this sense, variable selection is a central issue in clinical research. On the one hand, for statistical purposes, to overcome the curse of dimensionality ; and on the other hand, for practical purposes, to build transportable models, that could be applicable in a hospital environment. The literature gathers numerous methods of variable selection to address this issue. Without claiming to be exhaustive, we can cite to that extent the survival Lasso of Tibshirani [Tib97], the priority-Lasso of Klau *et al.* [KJH+18] which integrates a priori information on the relevance of the different omics measures, or the boosting approaches of Buehlmann [Bue06], Tutz and Binder [TB06]. However, due to the very high-dimensionality of the processed data, methods based mainly on the minimization of the fitting error will not always lead to the selection of the real model support. Indeed, the higher the number of predictors, the higher is the probability to find a random combination of variables minimizing the fitting error. Moreover, each of the existing methods deals with different approaches, and lead to different selections, sometimes leaving the practitioner undecided.

In this chapter, we propose a stepwise variable selection algorithm (SVSSA), suitable for multi-omics analysis. The idea is to take advantage of the performances of variable selection methods praised by the literature, by offering a consensus between the different approaches. As can be seen in Figure 5.1, our procedure is decomposed into five steps. Step 0 is optional and consists in removing outliers form the dataset. Steps 1 to 3, focus on reducing the size of the omics matrices, to bring them back to an order of magnitude close to that of the clinical data. It is only in step 4 that clinical data enters the selection process.

Algorithm 1 and Figure 5.1 provide an overview of the selection procedure. Each step of the process involves scoring based on four methods. Step 0 of outliers detection assigns an atypicality score to each individual in the dataset. For this preliminary step, only clinical information is taken into account to calculate the scores. Step 1 performs a large marginal selection within the omics matrices. The idea is to reduce, in the first instance, the size of the matrices to a reasonable level, to facilitate the execution of

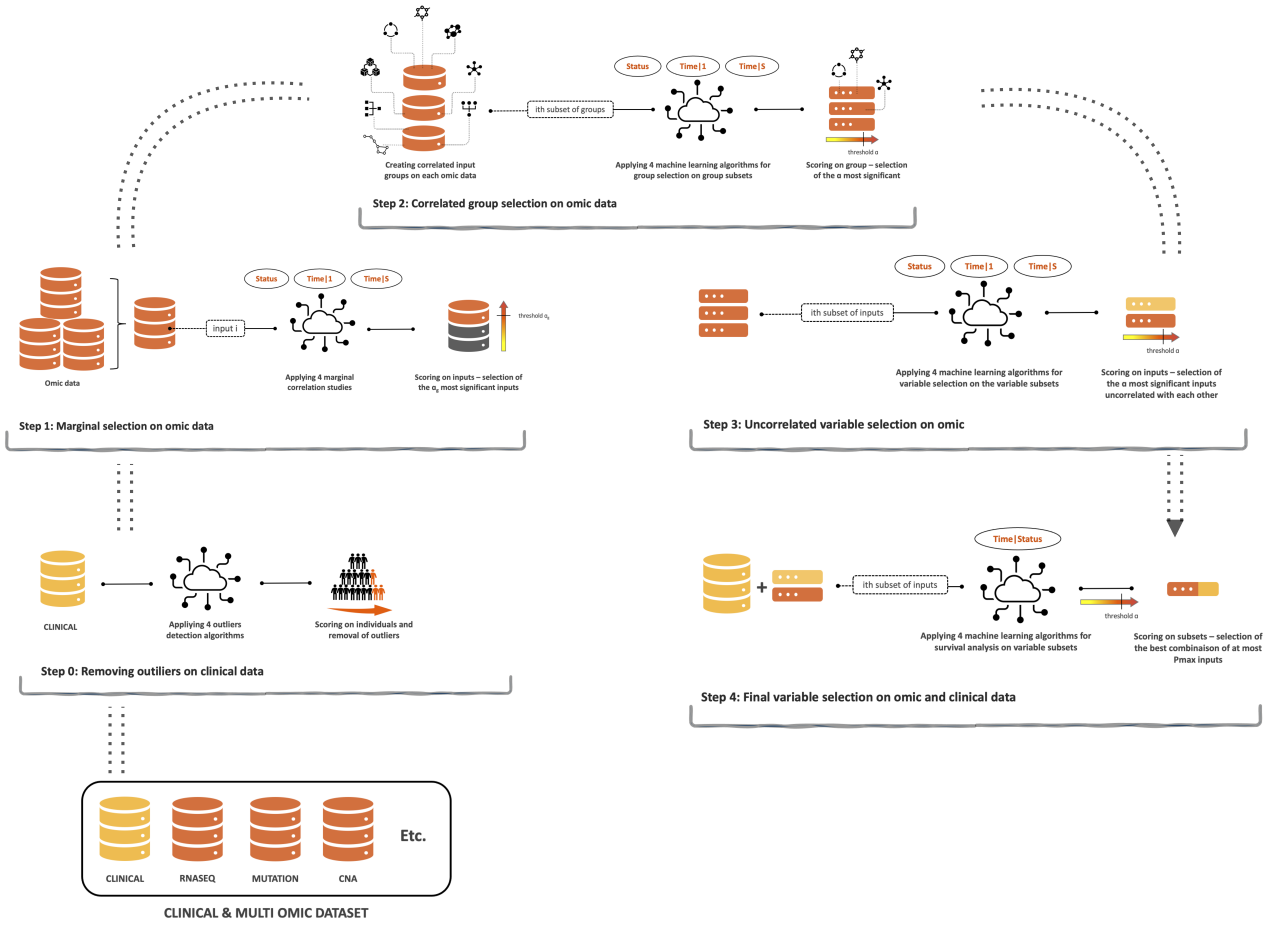


FIGURE 5.1 : Flowchart of the variable selection process using the SVSSA algorithm.

the following steps. In step 2, we exploit the correlation structures within the omics matrices by selecting groups of correlated variables. First, we build the different groups[1] with a clustering method, then we perform again a scoring at the group level. Step 3, in some way, consists in selecting the most relevant variables in each group, but indirectly. To do this, we perform a scoring at the variable level, then select the variables being at the same time the most relevant, but also the least correlated between them. Thus, we are seeking for a significant selection, while limiting information redundancy. Finally, in step 4 we make the final selection of at most $P^\star$ variables. This step combines a scoring procedure with a forward-backward selection procedure. The main idea of this whole process is to gradually reduce the size of the data, ensuring that at each step a consensus is reached on the choice of the selected variables. Once the selection by SVSSA is done, the practitioner can perform predictions by building the model of their choice with the selected variables.

The following sections present the different steps of SVSSA. We will not go into the details of how the different methods used work, but we refer the interested reader to the Appendix A, where we provide a concise presentation of the tools, scoring procedures and algorithms used in SVSSA, and for more details

1. or rather subgroups since the omics matrices themselves are considered as the main groups

please refer to the seminal articles. Moreover, note that this procedure does not aim to build a model but to select variables. Therefore, in some borrowed selection algorithms, we set aside the search for optimal penalization parameters, and set default parameters inducing sparsity. These default parameters can be modified by the practitioner. We make available the R codes of the SVSSA algorithm on Github[2].

Algorithm 1 SVSSA

Input: Dataset $Data = (\mathbf{Y}, \delta, \mathbf{X})$, group index $GId = (Id_1, ..., Id_m)$ where Id_1 are the indexes of the clinical variables if included, ultimate number of variables to select $P^\star$,

Optional: indicate whether the step$_0$ should be launched Step$_0$ = true, all optional inputs of Algorithms 3 to 6.

Output: Final dataset $Data_{\text{final}}$, variables selected $Selected\_vars$.

if Step$_0$ = true **then**

Step$_0$: run Algorithm 2 to remove outliers

end if

Step$_1$: run Algorithm 3 to perform a marginal selection on the omics matrices

Step$_2$: run Algorithm 4 to first build subgroups of variables, then perform group selection on the omics matrices

Step$_3$: run Algorithm 5 to perform a variable selection on the omics matrices

Step$_4$: run Algorithm 6 for the final selection on clinical data and the remaining omics variables

Now and for the rest of the chapter, consider that we have a sample of n independent observations (Y_i, δ_i, X_i), where for an individual i, Y_i corresponds to the observed survival time, δ_i to the censoring indicator, and X_i to the p-dimensional vector of the values taken on the predictors. Moreover, since the predictors are ordered in m groups of sizes $\kappa_1 + ... + \kappa_m = p$, X_i admits the decomposition $(\underline{\mathrm{X}}_{1\,i}, ..., \underline{\mathrm{X}}_{m\,i})$. The observed data are in $\mathbf{X} \in \mathbb{R}^{n \times p}$ where the i-th row is X_i^t, $\delta \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^n$, and we note $(\mathbf{X}^{(d)}, \mathbf{Y}^{(d)})$ the subsample containing only the individuals who observed the event. Note in particular that in all estimation phases where the target variable is only the survival time Y, the estimation is based on the subsample $(\mathbf{X}^{(d)}, \mathbf{Y}^{(d)})$.

5.1 Preliminary step: Removing outliers

Outlier detection is an important step prior to building any learning model. In survival analysis, it enables the detection of extreme individuals with a survival time that is too long or too short given their individual characteristics, and who may therefore degrade the model estimation. This preliminary step is optional, but highly recommended. It can be done manually by the practitioner or automatically when launching SVSSA. The procedure we present is based on the paper of Carrasquinha *et al.* [CVV18]. The authors proposed an anomaly scoring algorithm that performs a consensus between ten outlier detection methods on each individual of the dataset. We use their tools here, but we limit our procedure to four methods for computation time matter. The tools used here can be grouped into two approaches. First, the residual approach which represents one of the fundamental procedures in outlier detection. It is generally described

2. `https://github.com/EuniceOkome/SVSSA`

as the study of differences between the observed values of a variable of interest and the values predicted by a given regression model ; outliers are then represented by significant differences. As it stands, this definition is not perfectly applicable to survival models because it does not take censoring into account. We use three procedures that have been able to adapt to this constraint, thus improving the interpretability of the residual approach in the context of our study. The second approach is the predictive performance approach. The simple and intuitive idea is to penalize individuals who reveal a negative impact on the model predictive performances. This step relays only on the clinical data ; Algorithm 2 illustrates its flow. The methods used to measure the outlierness are the martingale residuals of Therneau *et al.* [TGF90], the deviance residuals of Therneau *et al.* [TGF90], the censored quantile regression residuals of Eo *et al.* [EHC14], and the Dual Bootstraps Hypothesis Testing of Pinto *et al.* [PCV15a]. The scoring procedure is based on the Rank Product test as suggested by [CVV18].

Algorithm 2 Removing outliers

Input: Dataset $Data = (\mathbf{Y}, \delta, \mathbf{X})$, group index $GId = (Id_1, ..., Id_m)$ where Id_1 are the indexes of the clinical variables if included,
Optional: rate of individuals to be removed $r = 0.01$.

Output: New dataset $Data_{s1}$.

$Data_{\text{used}}$ = clinical data
Compute the martingale residuals of the Cox model
Compute the deviance residuals of the Cox model
Compute the residuals of the censored quantile regression
Compute the Dual Bootstraps Hypothesis Testing
Perform the scoring using the Rank Product test
Remove the rn most atypical individuals from the dataset

5.2 Step 1: Marginal selection

The marginal selection step consists in an aggregation of four correlation analysis methods. We exploit the idea that a variable marginally correlated with the target variables is likely to have a significant impact on the final model. To compare the degrees of marginal correlations, we assign a significance score to each variable according to its level of association with the variables Y, δ, and the pair (Y, δ). The Algorithm 3 illustrates the flow of this selection step.

As discussed in the introduction section, this step does not include clinical predictors. The methods used to measure the significance of the correlation are the Spearman correlation test [Spe04], the Wilcoxon–Mann–Whitney test [Wil45, MW47], the Uno's concordance index (c-index) [UCP+11] and the integrated Brier score (iBrier) [GSSS99]. The scoring procedure is again based on the Rank Product test. Moreover, in order to exploit the multi-omics aspect of the data in the following steps, we want to ensure that each group has a nonzero size after the current selection step. To do so, we perform the scoring procedure within each omic group g and select the $\min\left[\, p_g,\, 0.1 \min_{g \in \{2,\dots,m\}} (p_g) + P^\star \,\right]$ most significant variables. This default setting aim to reduce all omics matrices to the same size, but the practitioner can assign different selection rates to each matrix.

Algorithm 3 Marginal selection

Input: Dataset $Data = (\mathbf{Y}, \delta, \mathbf{X})$, group index $GId = (Id_1, ..., Id_m)$ where Id_1 are the indexes of the clinical variables if included, ultimate number of variables to select $P^\star$,
Optional: number of variables to be select in each omic group $P_g^{(1)} = \min\left[p_g,\, 0.1 \min_{g\in\{2,\dots,m\}}(p_g) + P^\star\right]$.

Output: New dataset $Data_{s2}$, variables selected $Selected\_vars$.

```
for g = 2, m do
    for j ∈ Id_g do
        Compute the Spearman correlation test between Y^(d) and X^(d)_.j, and take the p-value
        Compute the W-M-W test between δ and X_.j, and take the p-value
        Measure the c-index of the Cox model based only on the variable j
        Measure the iBrier score of the Cox model based only on the variable j
    end for
    Perform the scoring using the Rank Product test
    Select the P_g^(1) most significant variables of this omic matrix
end for
```

5.3 Step 2: Selection of correlated groups

We have previously discussed the existence of correlation structures within omics data. In particular, within each grouping by omic type, we can observe more or less strong correlations between predictors. Co-regulated gene modules and gene regulatory networks included in RNA-sequencing data are a good illustration of this point. In this step, we want to exploit the fact that some variables have a joint expression on the model, and offer better results when their correlation structure is taking into account. To do so, we perform a selection by groups of correlated variables, segmented in two phases. First we build the groups using hierarchical clustering, then we aggregate four selection methods per group in order to assign a significant score to each of them. As in step 1, this procedure does not include the clinical data, and the selection models will be fit on the variables Y, δ and the pair (Y, δ). For a better understanding, let's break down the Algorithm 4 which summarizes the flow of this selection step.

Recall that the input dataset results from step 1. The clustering procedure is applied to each distinct omic group, and the number of clusters considered is proportional to the size of the initial group. This arbitrary parameterization is set to ceiling($p_g/50$), where p_g is the number of variables in the group g. The idea is to create clusters of reasonable sizes to allow the estimation algorithms to run smoothly. Furthermore, note that the initial clustering may lead to the isolation of some variables. In this case, the algorithm will progressively reduce the number of clusters until we obtain subgroups of two predictors minimum. For the sake of clarity, we now call *groups* the omic groups and *subgroups* the clusters resulting from the hierarchical clustering. In the estimation phase, the algorithm applies four group selection methods on three subgroups of variables. Each subset of subgroups is constructed randomly, so that any subgroup is included in a single subset. In order to test different combinations of subgroups and obtain relevant results, we iterate this estimation phase five times. Once the estimations are completed, each subgroup is assigned a significant score. The most significant are then selected so as to keep approximately $\min(\,0.1p + P^\star,\, p\,)$ variables. The method used in the estimation procedure are our Bayesian PGGM (see Chapter 3), the logistic group Lasso of Meier *et al.* [MVDGB08], the IPF-Lasso of Boulesteix *et al.* [BDBJF17], and the Block Forest of Hornung and Wright [HW19]. Finally, the global scoring consists simply in counting the

Algorithm 4 Correlated variable groups selection

Input: Dataset $Data_{s2} = (\mathbf{Y}, \delta, \mathbf{X})$, group index $GId = (Id_1, ..., Id_m)$ where Id_1 are the indexes of the clinical predictors if included, ultimate number of variables to selected $P^\star$,
Optional: number of iterations for the estimation procedure $nIt = 5$, number of subgroups in each model $ngr = 3$, model parameters, rate of subgroups to be selected $P^{(2)} = \min(\, 0.1p + P^\star, \, p\,)$.
Output: New dataset $Data_{s3}$, variables selected $Selected\_vars$
for $g = 2, ..., m$ **do**
p_g = number of variables in the group g
$\text{subGId}_g = (Id_{g1}, ..., Id_{gm})$ subgoup index obtained by hierarchical clustering on g
end for
$\text{subGid} = \bigcup_{g=2,...,m} \text{subGid}_g$
for $it = 1, nIt$ **do**
Create random subsets of ngr subgroups in subGId, s.t. each subgroup is included in one subset
for each subset **do**
Compute a bayesian PGGM model and give +1 to the score of subgroups selected
Compute a group Lasso model and give +1 to the score of subgroups selected
Compute an ipf-Lasso model and give +1 to the score of subgroups selected
Compute a Block Forest model give +1 to the score of subgroups verifying $w_{sgj} > 0.75$
end for
end for
Select the $P^{(2)}$ of the subgroups with the highest score

number of times that a given subgroup has been selected by the different methods.

5.4 Step 3: Selection of decorrelated variables

The main goal of step 2 was not only the selection of correlated groups, but also a selection of relevant variables based on random patterns favorable to their expression, *i.e.* combined with variables that should highlight their importance through joint actions. With the resulting data as support, we now want to recover variables that are both significant and as uncorrelated as possible, in order to limit information redundancy. To do so, we follow a reasoning similar to De Jay *et al.* in their mRMRe method [DJPCO+13], but adapting it to our consensual context. For a better understanding, let's break down the Algorithm 5 which summarizes the flow of this third selection step.

First, let's denote $p^{(o)}$ the number of non-clinical predictors in $Data_{s3}$, and let $nvar = p^{(o)}/10$ and $nrep = p^{(o)}$ be the default parameters. This step is divided into two phases, an estimation phase and a selection phase. In the first one, the algorithm begins by constructing subsets of $Data_{s3}$ each containing $nvar$ of randomly drawn non-clinical predictors, so that any variable is included in a single subset. Then, the estimation phase goes on to apply four variable selection methods on these samples. In order to test different combinations of variables and obtain a relevant selection, this phase is iterated $nrep$ times. The method used in the estimation procedure are our Structural PGGM (see Chapter 2), the logistic Lasso [Lok99, SK03], the Gradient boosting with component-wise for Cox model [Bue06, DB16], and the Spike-and-Slab Lasso Cox of Tang *et al.* [TSZY17]. As in the previous step, the results of this estimation phase provide a significance score for each variable, expressed as the number of times it was selected by the different methods. However, we will not only select the variables with the highest scores, but those

Algorithm 5 Uncorrelated variables selection

Input: Dataset $Data_{s3} = (\mathbf{Y}, \delta, \mathbf{X})$, group index $GId = (Id_1, ..., Id_m)$ where Id_1 are the indexes of the clinical predictors if included, list of remaining subgroups from step 2 subGroups, ultimate number of variables to select $P^\star$,
Optional: number of variables in each model $nvar = p^{(o)}/10$ where $p^{(o)}$ is the number of remaining non-clinical predictors, number of iterations $nrep = 10\,nvar$, rate of variables to be selected $\tau = 0.5$, model parameters
Output: New dataset $Data_{s4}$, variables selected $Selected\_vars$
  for $it = 1, nrep$ **do**
    Create random subsets of $Data_{s2}$ containing 10% of non-clinical variables, s.t. each variable is included in one subset
    for each subset **do**
      Compute a structural PGGM and give +1 to the score of variables selected
      Compute a logistic Lasso and give +1 to the score of variables selected
      Compute a bmlasso and give +1 to the score of variables selected
      Compute a glmboost and give +1 to the score of variables selected
    end for
  end for
  $nsel =$ ceiling$(\tau\, p^{(o)})$
  $Selected\_vars =$ the variable j with the highest score
  while number of predictor selected $< nsel$ **do**
    $score_j^{(p)} = score_j$ penalized by the correlation with the selected predictors
    Add to $Selected\_vars$ the variable j with the highest penalized score
  end while

that also have a low level of correlation with the other selected variables.

5.5 Step 4: Final selection

Throughout the previous steps, we refined the non-clinical dataset to keep only the most relevant variables. We can now integrate the clinical data into the selection process, without concern about being unfairly underestimated due to their small size. This last step focuses on the estimation of survival models, and is composed of three phases ; an estimation phase, a first selection phase by scoring, and a second selection phase with forward-backward approach. The Algorithm 6 summarizes its flow.
Adding clinical data forces us to reflect on an additional problem of clinical analysis ; in particular, according to practitioners, we often observe a *poison* effect of certain omics variables on clinical variables. In other words, the presence of certain variables in a model can negatively impact the evaluation of some clinical variables. To limit this inconvenience, we segment this last estimation phase in three parts. First, we build the models using each individual group separately, then by pairs of groups, and finally we consider the interaction of all groups together. In this way, we can assess the relevance of the variables in different contexts. The method used in the estimation procedure are the sparge-group Lasso of Simon *et al.* [SFHT13], the priority-Lasso of Klau *et al.* [KJH+18], the random survival Forest of Ishwaran *et al.* [IKBL08] and the Likelihood-based boosting for Cox model [TB06, BASB09, DB16]. The first phase of selection consists simply in selecting the $P^\star$ variables with the highest score. The second selection phase aims to find the best performing combination of variables, among the most significant ones. To this end,

Algorithm 6 Final variables selection

Input: Dataset $Data_{s4} = (\mathbf{Y}, \delta, \mathbf{X})$, group index $GId_{s2} = (Id_1, ..., Id_m)$ where Id_1 are the indexes of the clinical predictors if included, ultimate number of variables to be selected $P^\star$,
Optional: model parameters.
Output: New dataset $Data\_final$, variables selected $Selected\_vars$.
for each group, each pair of groups and all groups together **do**
Compute a sparse-group Lasso and give +1 to the score of variables selected
Compute a priority-Lasso favoring clinical data and give +1 to the score of variables selected
Compute a survival Forest and give +1 to the score of variables selected
Compute a Coxboost and give +1 to the score of variables selected
end for
Select the $P^\star$ variables with the highest score
Select the best combination of variables using a forward-backward approach that minimize the iBrier score on the learning dataset

we perform a forward-backward selection procedure in order to minimize the cross-validation iBrier score.

5.6 Summary

We end this presentation of the SVSSA method with the Table 5.1 that provides a summary of the methods and algorithms used.

5.7 Application on real data

In this section, we explore the problem that motivated this thesis, the selection of relevant features for the diagnosis of TNBC. As presented in the Introduction, TNBC is a particularly challenging tumor form in oncology. Because of the difficulties encountered by oncologists when evaluating the diagnosis, there is a significant relapse rate in patients and a rapid progression of the disease to a nearly incurable metastatic stage. The SVSSA method seeks to overcome these limitations by providing a reasonable amount of significant prognostic factors, thus impacting the therapeutic follow-up of patients. In order to evaluate its relevance in this context, we will compare its predictive performances with those of other variable selection methods, on a real dataset.

5.7.1 Presentation of the dataset

Our study is based on data from the multi-omics TNBC cohort at Fudan University Shanghai Cancer Center (FUSCC), available on the National Omics Data Encyclopedia (NODE) platform[3]. This cohort included 465 TNBC patients, for whom we have one to three matrices of omic data. In order to perform a multi-omics analysis, we restricted our study to patients displaying the three omic measures. This left us with 233 patients and 253 411 explanatory variables distributed in four matrices: clinical, RNA sequencing (RNAseq), mutation and copy-number alteration (CNA). We are clearly in a high dimensional framework and more precisely in multi-omics data analysis. Figure 5.2 shows the distribution of the omics

3. NODE: `https://www.biosino.org/node/project/detail/OEP000155`

Step	Approach	Learner	Method	Package::function	Tuning	Selection
Step 0	Residual	martingale	Martingale residuals	stats::resid	-	outliers
	Residual	deviance	Deviance residuals	stats::resid	-	outliers
	Residual	quantile residuals	Quantile regression residuals	OutlierDC::odc	-	outliers
	Performance	DBHT	Dual Bootstraps Hypothesis Testing	OutlierDC::odc	-	outliers
Step 1	Correlation	correlation	Spearman correlation	stats::cor.test	-	-
	Correlation	W-M-W	Wilcoxon–Mann–Whitney test	stats::wilcox.test	-	-
	Performance	c-index	Uno's concordance index	survAUC::UnoC	-	-
	Performance	iBrier	Integrated Brier score	survAUC::predErr	-	-
Step 2	Clustering	RF	Unsupervised Random Forest	randomForest::randomForest	OOB	-
	Clustering	HAC	Ward hierarchical clustering	stats::hclust	-	-
	Bayesian	bPGGM	Bayesian PGGM	GibbsPGGM	sparsity	groups
	Groups	ggLasso	Logistic group Lasso	gglasso::gglasso	sparsity	groups
	Adaptative	ipfLasso*	IPF-Lasso	ipflasso::cvr.ipflasso	CV	variables
	Forest	blockForest*	Block Forest	blockForest::blockfor	OOB	Groups
Step 3	Structural	sPGGM	Structural PGGM	Estim	sparsity	variables
	-	lasso	Logistic Lasso	glmnet::glmnet	sparsity	variables
	Boosting	gboost	Gradient boosting	mboost::glmboost	sparsity	variables
	Bayesian	bmlasso	Spike-and-Slab Lasso Cox	BhGLM::bmlasso	sparsity	variables
Step 4	Groups	SGL	sparse-group Lasso	SGL::SGL	sparsity	variables
	Favoring	priorLasso*	priority-Lasso	priortiylasso::priorityllasso	CV	variables
	Forest	ranger	random survival Forest	randomForestSRC::rfsrc	OOB	variables
	Boosting	cboost	Likelihood-based boosting	CoxBoost::cv.CoxBoost	sparsity	variables

TABLE 5.1 : Summary of the methods used in SVSSA.

matrices in the dataset. We can see that the size of the clinical data is derisory compared to that of the other measures (0.02 % of the dataset), while by contrast the CNA matrix is predominant (85.87 % of the dataset). This distribution illustrates the need for late selection within the clinical data so that their under-representation does not lead to an underestimation of their prognostic relevance.

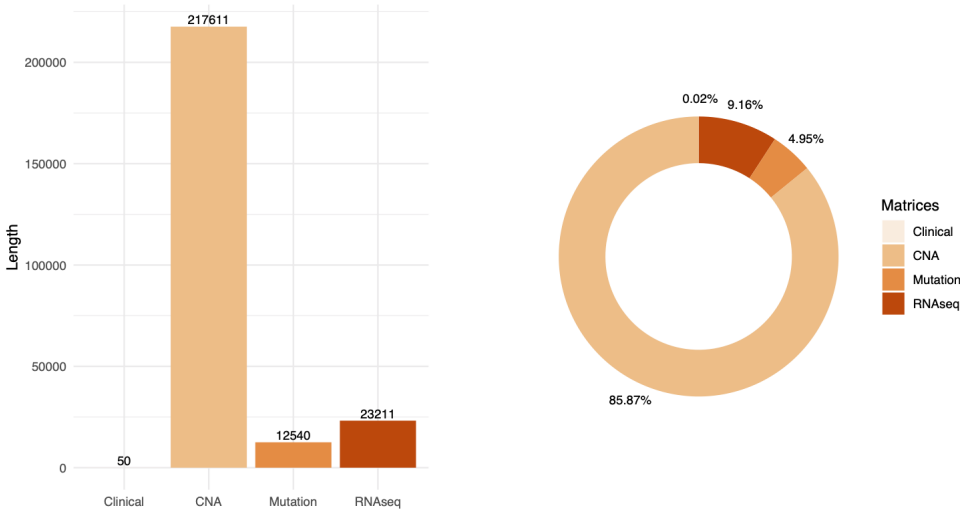


FIGURE 5.2 : Distribution of omics matrices in the dataset.

Before performing the analysis, pre-processing of the dataset is necessary due to the specificities of clinical data. Firstly, in cohort data, there is generally a lot of missing data, particularly because the individuals recruited do not wish to carry out all the planned tests. Furthermore, the variables considered in the clinical matrices are presented in different formats. In order for the methods selected in this analysis to work properly, it is necessary to remove the missing information from the dataset and to convert the variables that are not already in numerical format.

The treatment of missing data must be done with caution so as not to lose too much clinical information. In our dataset, we have 15 variables and 233 observations with at least one missing information. Since no observation in the sample is complete, it is necessary to study the a priori relevance of each variable, in order to find a balance in the removal of both variables and individuals, and thus maintain a suitable amount of information. We therefore first removed 18 irrelevant variables, some of which contained missing data. Table 5.2 lists these variables and the reasons for their removal. We also removed patient *FUSCCTNBC389* because she had zero follow-up time, and therefore represents an outlier. At this stage of the study, the clinical matrix has 232 observations and 32 variables, for which we have 122 observations and 11 variables with at least one missing value. The boxplots on the left side of Figure 5.3 show the distribution of missing data within these incomplete variables and observations. Since the median of the variable was 25, we decided to remove the variables with at least 25 missing data so as not to further reduce the sample size. After that, only three observations had missing data, we removed them from the dataset as well. To summarize, in addition to the variables in Table 5.2 and patient *FUSCCTNBC389*, we removed the variables *SNF_Subtype*, *Mutation_Subtype*, *BRCA1.2.MUT.20160718*, *sTILs*, *iTILs*, *Grade* and the patients *FUSCCTNBC134*, *FUSCCTNBC170*, *FUSCCTNBC245*.

Variables	Reason for removal
Exome_sequencing	indicates whether the patient has received exome sequencing, *i.e.* whether we observe her values in the Mutation matrix. This is the case for all patients in our dataset, so this variable is useless for model building.
RNA_sequencing	indicates whether the patient has received RNA sequencing, *i.e.* whether we observe her values in the RNAseq matrix. This is the case for all patients in our dataset, so this variable is useless for model building.
OncoScan_Array	indicates whether the patient has been subjected to whole genome copy number assay, *i.e.* whether we observe her values in the CNA matrix. This is the case for all patients in our dataset, so this variable is useless for model building.
Sex	indicates the sex of the patient. Since all patients are women, this variable is useless for model building.
BRCA1.MUT.20160718	indicates whether the patient underwent a BRCA1 mutation prior to 7/18/2016.This variable provides redundant information with the variable *BRCA1.2.MUT.20160718* which indicates whether the patient underwent a BRCA1 or BRCA2 mutation prior to 7/18/2016.[a]
BRCA2.MUT.20160718	same reason as *BRCA1.MUT.20160718*.
DNA_QC_Failed	indicates whether the DNA quality control has failed. This variable takes the value *FALSE* for all patients, and is useless for model building.
HRD	indicates the estimation of homologous recombination deficiency score. It was calculated as the sum of three scores: telomeric allelic imbalance (NtAI), loss of heterozygosity (LOH) and large scale transition (LST) [JMS+19]. This variable provides redundant information with the variables related to the three scores.
Histology_extended	contains 218 missing values.
RFS_time_Days	indicates the observed survival time of the patient in days. This variable provides redundant information with the variable *RFS_time_Months* which indicates the observed survival time in months.
Followup_Month	indicates the time the patient has spent in the project in months. It was calculated as the difference between the date of last follow-up and the date of surgery (which is equivalent to date of inclusion). This variable provides redundant information with the variable *RFS_time_Months*. In particular, they have the same value for censored observations.
Date_of_surgery	same reason as *Followup_Month*.
Date_of_last_followup	same reason as *Followup_Month*.
No_Chemotherapy	indicates whether the patient has not received chemotherapy. This variable provides redundant information with the variable *Chemotherapy* which indicates whether the patient has received chemotherapy.
ER_IHC_score	indicates the result of the immunohistochemistry test that determines the presence of estrogen receptor in the patient's cancer cells. This variable takes the value *negative* for all patients since it is a particularity of the TNBC[b]. It is therefore useless for model building.
PR_IHC_score	same reason as *ER_IHC_score* but with progesterone receptor.
ERBB2_IHC_score	same reason as *ER_IHC_score* but with HER2 protein.
ERBB2_FISH	contains 168 missing values.

TABLE 5.2 : Presentation of the first clinical variables removed and the reasons behind these choices.

a. The pathogenic mutation of the BRCA1 and BRCA2 genes affects a minority of patients with breast cancer, but its impact is major. It is one of the most powerful risk factors for this pathology, since 65% and 45% of the women affected, respectively, develop the disease [APN+03].

b. The term triple-negative breast cancer refers to the fact that cancer cells test presence-negative for estrogen receptors, progesterone receptors and the HER2 protein [Soc].

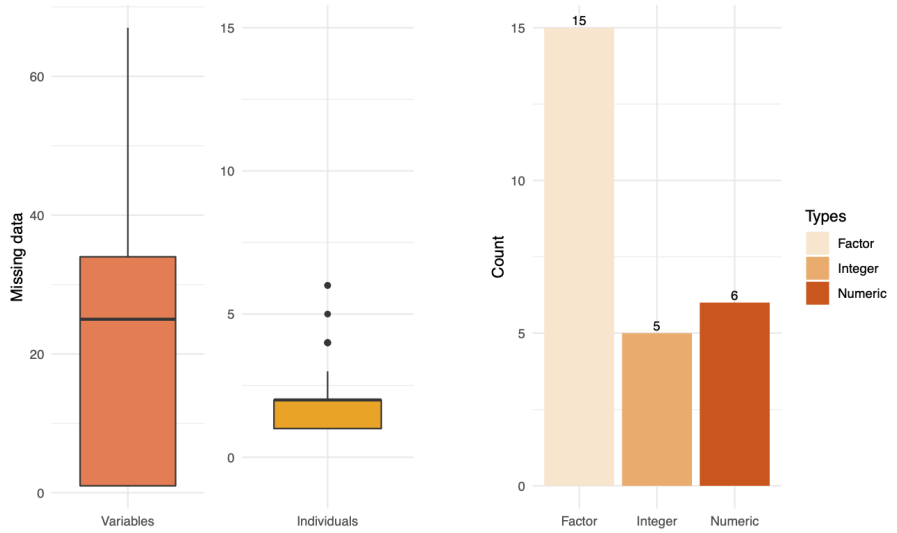


FIGURE 5.3 : Distribution of missing data restricted to variables and observations with at least one missing data after the first removal (left), and distribution of different variable formats in the clinical matrix after all missing data has been removed (right).

After removing the missing data, the clinical matrix is now composed of 229 observations and 26 variables of different formats. The bar chart on the right of Figure 5.3 shows the distribution of these formats within the matrix. For the proper functioning of the algorithms we transform the categorical variables into vectors of quantitative variables, using the One-Hot encoding method [Bro22]. After the conversion, the clinical matrix has 65 variables, including 2 variables of interest (observed survival time and survival status) and 63 explanatory variables.

5.7.2 Presentation of the algorithms

In this study, we will compare the performance of our selection algorithm with nine methods that have been proven to be effective in the literature. However, since the SVSSA method has mainly a selection objective and not a prediction one. We will focus on the performance in terms of variable selection, in order to compare our approach to other methods. To do so, we use each method to perform a selection of variables assumed to be significant, and then we apply a Cox model with the selected variables to perform predictions. The choice of methods considered is mainly based on the benchmark study by Herrmann *et al.* [HPH+21] on survival predictions in multi-omics analysis. One can also notice that most of these comparative methods play a role in the SVSSA approach. This will allow to juxtapose the respective performances of these methods with their joint performances. Table 5.3 provides an overview of the methods considered. Five approaches are covered, distributing the methods as follows:

– **the penalized regression approach:** survival Lasso [Tib97], IPF-Lasso [BDBJF17], priority-Lasso [KJH+18], GRridge [VDWLV+16], sparse-group Lasso [SFHT13],

– **the boosting approach:** gradient boosting [Bue06, DB16], Likelihood-based boosting [TB06, BASB09, DB16],

– **the nonlinear approach by decision trees:** survival random Forest [IKBL08],

– **the bayesian approach:** spike-and-slab Lasso Cox [TSZY17],

– **the aggregation approach:** Stepwise Variable Selection for Survival Analysis.

Learner	**Method**	`Package::function`	**Tuning**
lasso	survival Lasso	`glmnet::cv.glmnet`	10-fold CV
ipfLasso*	IPF-Lasso	`ipflasso::cvr.ipflasso`	10-fold CV
priorLasso*	priority-Lasso	`priortiylasso::prioritylasso`	10-fold CV
SGL*	sparse-group Lasso	`SGL::SGL`	sparsity
gboost	gradient boosting	`mboost::glmboost`	sparsity
cboost	likelihood-based boosting	`CoxBoost::cv.CoxBoost`	sparsity
ranger	survival random Forest	`randomForestSRC::rfsrc`	OOB
bmlasso	spike-and-slab Lasso Cox	`BhGLM::bmlasso`	sparsity
clinicalCox	Cox model	`survival::coxph`	No
SVSSA*	SVSSA	`SVSSA` and `survival::coxph`	sparsity

TABLE 5.3 : Summary of the methods used in the comparative study, the symbol * indicates the methods incorporating the group structure in their estimation procedure.

Each of these methods performs selection either at the variables level. The tuning of the parameters is carried out according to the desired degree of sparsity, when this option is proposed, otherwise it is done by cross-validation. As previously discussed, a limited number of variables is strongly recommended to ensure the transportability of the models obtained in a clinical setting. Therefore, we have set the maximum number of variables to be selected at 100. Finally, we include in the comparative study a Cox model built with only clinical data. This will serve as a baseline model to evaluate the contribution of omics data in the context of our study.

5.7.3 Variable selection with the SVSSA method

In this section, we present the results obtained by SVSSA on the whole dataset. As mentioned above, the maximum number of predictors to be reached is 100. The input parameters are the default parameters unless otherwise specified.

Preliminary step: Removing outliers

We start with the preliminary step of outlier detection. During this step, the algorithm designated 10 individuals as outliers. Figure 5.4 illustrates, on the one hand the intersections between the different methods used for sets of 20 most atypical individuals according to each measure ; and on the other hand the proportion, for each method, of individuals designated by SVSSA present in its list of the 20 most atypical. First notice on the left graph that only one individual is considered as outlier by the four

methods, and five other individuals are considered as outliers by three of them. These first 6 individuals were considered as significantly atypical by SVSSA. Moreover, the bar chart shows that SVSSA aligned mainly with the martingale residuals approach since all the detected individuals were also considered as the most atypical by this method. The DBHT method on the other hand showed less impact on this example. This last point can also be explained by the fact that it is the only approach based on predictive performance, and therefore it offers less common results with the other methods.

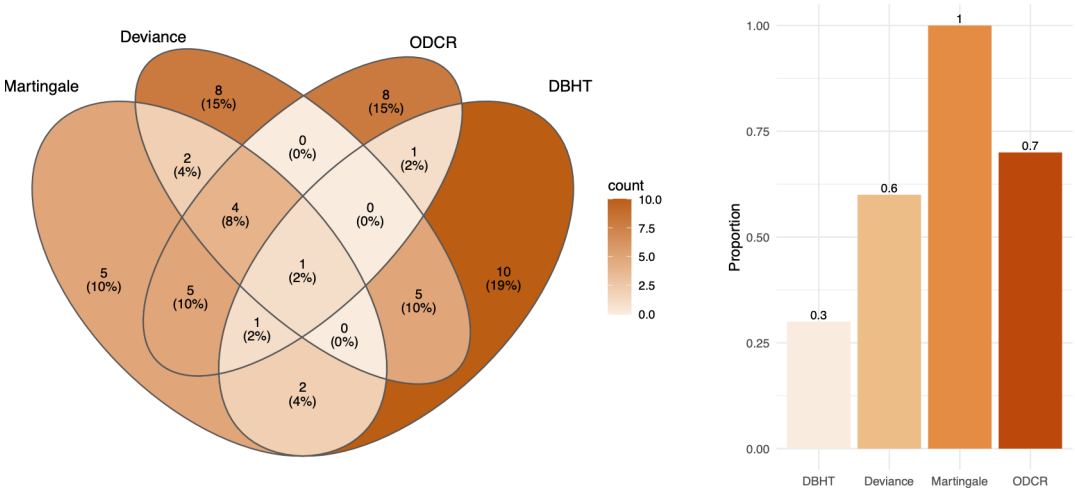


FIGURE 5.4 : Step 0: intersections between the four methods used for sets of 20 most atypical individuals according to each measure (left), proportion of outliers designated by SVSSA in each list of 20 most atypical individuals (right).

Step 1: Marginal selection

After removing the outliers from the dataset, we proceed to the variable selection. For the first step of marginal selection, we set the selection rates to 10% for the RNAseq and mutation matrices, and 1% for the CNA matrix. By applying a stricter selection on the CNA matrix, we return to comparable scales for the three matrices, and guard against the bias that would have been caused by the overrepresentation of the CNA matrix in the models. Thus, by SVSSA, we selected 2 421 RNAseq variables, 1 354 mutation variables and 2276 CNA variables ; for a total of 6 112 explanatory variables (including clinical ones). Figure 5.5 summarizes the results obtained in this step. The graph on the left illustrates the intersections between the methods for the same selection rate. We can notice that 155 variables were selected unanimously by the four methods, and 1 169 variables were selected by three of them. The bar chart on the right shows that the methods based on the c-index, the Brier score, and the Wilcoxon test had the most impact for this selection step. Indeed, the measure of correlation with survival time resulted in more isolated choices - 4898 variables were selected by this method alone - which were not considered as relevant by SVSSA.

Step 2: Selection of correlated groups

The dataset now consists of a reasonable amount of explanatory variables (6 112). Figure 5.6 represents a portion of the correlation plots of the different omics matrices. They clearly show the existence of group

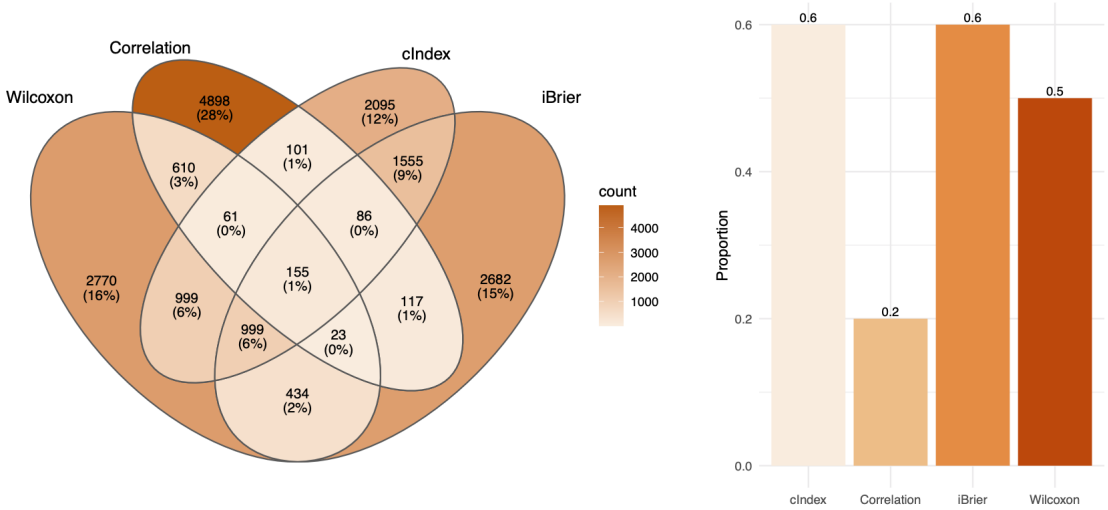


FIGURE 5.5 : Step 1: intersections of the sets of variables selected by each of the four methods (left), proportion of variables selected by SVSSA in the sets of variables selected by each individual method (right).

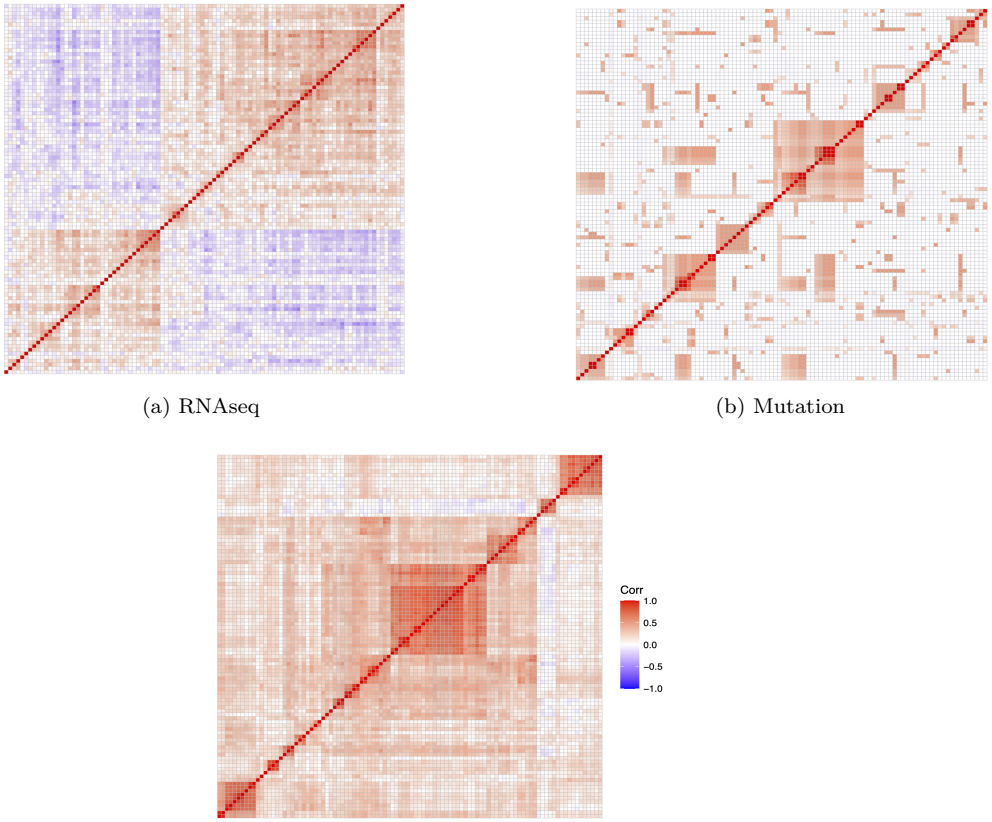
structure between variables, in particular within the RNAseq data, emphasizing the interest of this group selection step. For the first step of marginal selection, we set the selection rates to 30%. The clustering procedure resulted in the construction of 107 groups of variables: 33 RNAseq groups, 28 mutation groups, and 46 CNA groups. During the selection phase, 38 groups were considered relevant by SVSSA. This lead to a selection of 1 889 omics variables, and a total of 1 950 explanatory variables (including clinical ones). Figure 5.7 summarizes the results obtained in this step. We notice that only 2 groups were selected by the four methods, while 24 were selected by three of them. Finally, the bar chart shows that the Block Forest, Bayesian PGGM and group Lasso methods had an equivalent impact for this selection step, while the IPF-Lasso is slightly behind.

Step 3: Selection of decorrelated variables

Now we want to select a set of relevant variables that are as decorrelated as possible to minimize information redundancy. For this step we set the selection rate at 10%, that lead to a selection of 189 variables, for a total of 250 explanatory variables (including clinical ones). Figure 5.8 summarizes the results obtained in this step. The bmlasso method did not select any of the candidate variables, so it is excluded from the intersection plot, and naturally it is associated with a null proportion on the bar chart. Among the selected variables, 9 have been detected by the three remaining methods, and we also notice that most of the variables considered significant by SVSSA come from the gradient boosting selection.

Step 4: Final selection

For this last step, we include clinical data in the selection procedure, as the size of the omics matrices is reasonable enough not to worry about underestimating the clinical information. The maximum number of variables to select was set at 100. This setting lead to a selection of 19 explanatory variables. Figure 5.9



(a) RNAseq

(b) Mutation

(c) CNA

FIGURE 5.6 : Correlation plot of the first 100 variables for each comedy matrix

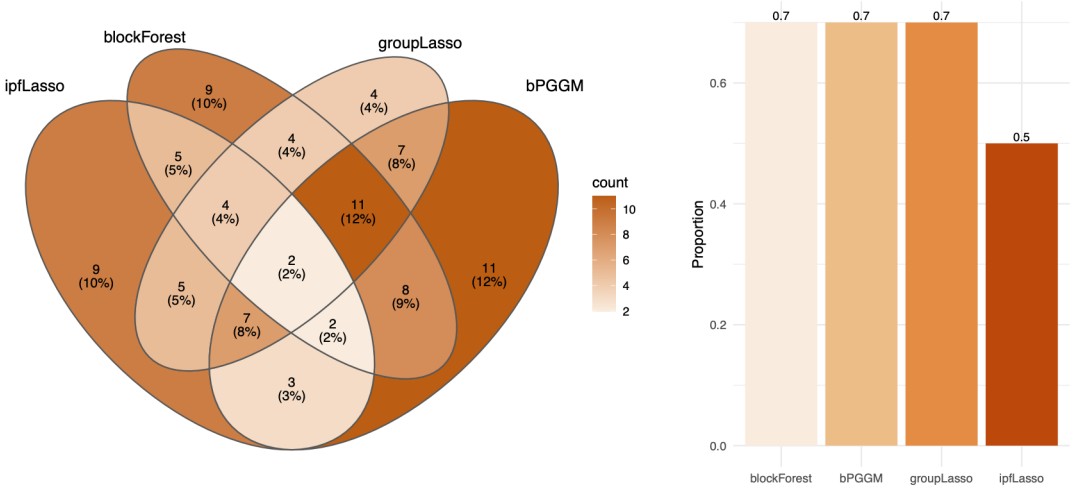


FIGURE 5.7 : Step 2: intersections of the sets of groups selected by each of the four methods (left), proportion of groups selected by SVSSA in the sets of groups selected by each individual method (right).

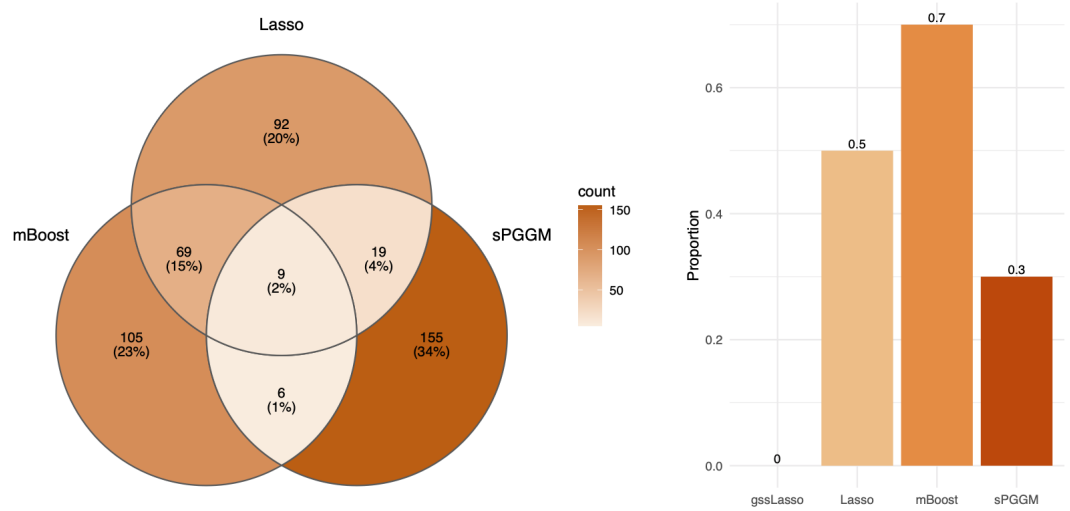


FIGURE 5.8 : Step 3: intersections of the sets of variables selected by each of three methods (left), bmlasso did not select any of the candidate variables ; proportion of groups selected by SVSSA in the sets of groups selected by each individual method (right).

summarizes the results obtained in this step. We can see that the selections obtained by the four methods are very homogeneous. Moreover although after the first selection phase, the four methods agreed on 83 variables, only 19 of them were kept after the forward-backward selection. Figure 5.10 shows the evolution

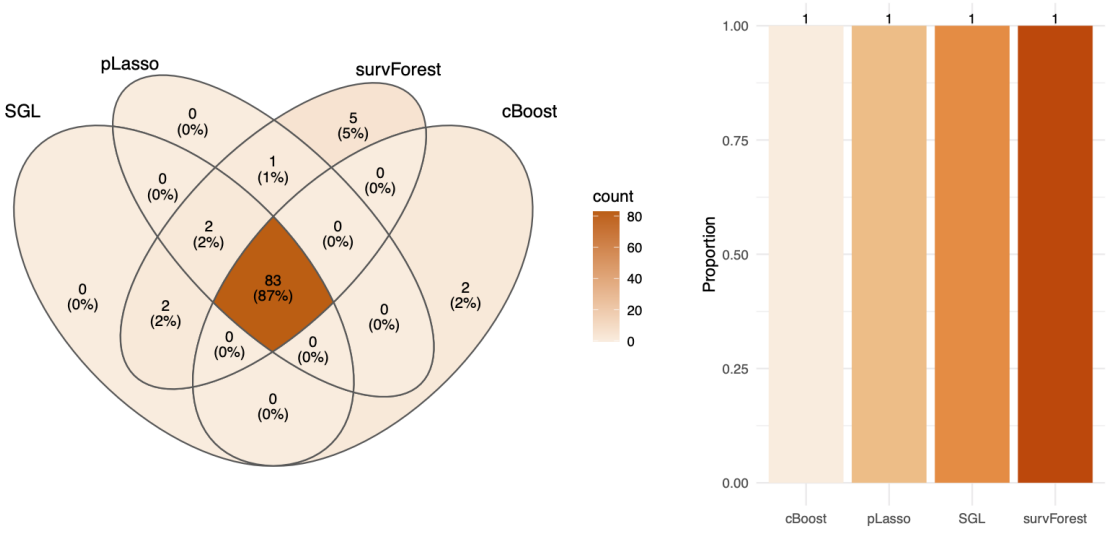


FIGURE 5.9 : Step 4: intersections of the sets of variables selected by each of the four methods (left) ; proportion of variables selected by SVSSA in the sets of groups selected by each individual method (right).

of the matrix sizes after each selection step. The mutation matrix was completely removed in step 2. This observation is in agreement with the medical literature.

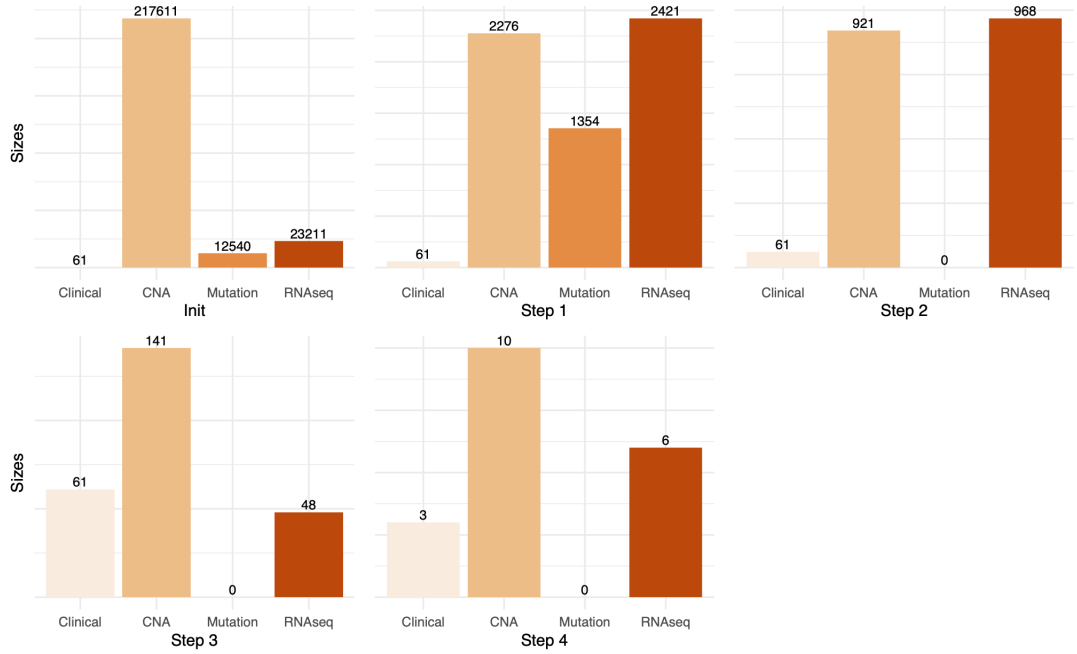


FIGURE 5.10 : Evolution of the size of the matrices after each selection step.

5.7.4 Performance comparison

As mentioned above, we want to compare the performance of SVSSA in terms of variable selection with nine other methods. Each of them is used only in the selection phase, the prediction phase being carried out by a Cox model. For all the comparative methods considered, we used the dataset resulting from the removal of outliers ; the initial dataset led to many model fitting failures. Three measures are considered: the Uno's concordance index and the AUC score for the discrimination power, and the iBrier score for the predictive power ; each of them was evaluated using the package `survAUC`. Finally, two types of accuracy are presented ; the training accuracy and the validation accuracy.

The training accuracy is evaluated through a 10-fold stratified cross-validation on the same dataset used for the selection procedure. For all the methods considered, we first perform a variable selection phase using the whole dataset, then we perform the cross-validation on Cox models built with the selected variables. Table 5.4 shows that SVSSA offers good performances in terms of discrimination power, due to its high c-index and AUC score. Similarly, if we look at the standard deviation values for these measures, we see that the results of SVSSA are quite homogeneous on the different cross-validation folds. Nevertheless, although SVSSA also has a very good predictive power, the priority-Lasso outperforms it on this criterion with a very low iBrier, associated with a low standard deviation as well. It is rare that a method offers the best performances over all criteria [HPH+21], hence the interest in measuring different ones to get an overview of its capabilities. This can be seen, for example, with the lasso and the bmlasso which minimize the iBrier quite well, but do not offer interesting results on the c-index and the AUC. Since SVSSA performs well on all three measures, this first analysis argues for a consistency of our procedure. However, it is necessary to verify this premature observation on a validation set independent

of the training set. The evaluation of the validation accuracy is a standard training-validation procedure.

Learner	c-index		iBrier		AUC score		selected
	Mean	sd	Mean	sd	Mean	sd	
lasso	0.797	0.204	0.091	0.034	0.814	0.212	5
ipfLasso	0.848	0.120	0.125	0.095	0.869	0.115	13
priorLasso	0.913	0.058	**0.083**	0.037	0.932	0.042	21
SGL	0.709	0.229	0.131	0.070	0.734	0.188	5
gboost	0.929	0.051	0.134	0.200	0.936	0.051	18
cboost	0.773	0.233	0.137	0.099	0.802	0.262	70
ranger	0.594	0.251	0.245	0.213	0.551	0.239	100
bmlasso	0.766	0.208	0.090	0.039	0.822	0.196	15
clinicalCox	0.683	0.238	0.202	0.196	0.727	0.187	-
SVSSA	**0.938**	0.066	0.095	0.103	**0.943**	0.058	19

TABLE 5.4 : Mean and standard deviation of performance measures obtained by 10-fold cross-validation on the training set.

The dataset is divided into two independent subsets, the training set represents 80% of the initial dataset, and the test set 20%. For each method, the selection phase is performed on the training set, and the Cox models are also built on this set restricted to the selected variables. The predictions are then applied on the validation set. The results obtained are reported in Table 5.5. While the ipf-Lasso, the priority-Lasso and SVSSA offer results close to those reported in Table 5.4, this is not the case for most methods. This observation argues for the existence of an overfitting bias when evaluating the training accuracy of these methods. Moreover, the observations previously made on SVSSA and the Lasso priority are again observed here. The performances of the two methods are quite close, and SVSSA outperforms the priority-Lasso on two of the three measures. Finally, we see that SVSSA offers consistent results on this dataset.

Computation time

Another important criterion for this comparative analysis is the computation time. SVSSA offers good performance in terms of variable selection, but this procedure is based on a superposition of several models, which makes it by construction slower than the comparative methods. We can see in Table 5.6 that SVSSA requires about 9 hours of additional compilation time to the priority-Lasso which also offers very good results. Note, however, that the computation time associated with SVSSA varies greatly depending on the settings of the optional parameters. A parameter that can be modified easily without really damaging the results is the number of iterations of step 2 (see nIt in Algorithm 4). By lowering it to 1 instead of 5 we can save a few hours of computation in step 2. Table 5.7 show the results of the two configurations in terms of training accuracy and computation time. The shorter configuration had better results, both compared to the initial SVSSA and to the priority-Lasso, but the computation time

Learner	**c-index**	**iBrier**	**AUC score**
lasso	0.576	0.094	0.655
ipfLasso	0.818	0.075	0.865
priorLasso	0.971	**0.046**	0.975
SGL	0.681	0.087	0.760
gboost	0.406	0.109	0.270
cboost	0.763	0.074	0.761
ranger	0.469	0.101	0.514
bmlasso	0.337	0.154	0.301
clinicalCox	0.885	0.051	0.914
SVSSA	**0.986**	0.050	**0.994**

TABLE 5.5 : Mean of performance measures obtained on the validation set.

is still significant. However, we must consider that this approach is intended for the analysis of clinical data which take several years to collect, and only require a variable selection phase to be exploited in a hospital environment. The contribution in performance and the serenity offered by a consensus selection compared to the depreciation in terms of compilation time is at the practitioner's discretion.

lasso	ipfLasso	priorLasso	SGL	gboost	cboost	ranger	bmlasso	SVSSA
0.038	1.224	2.448	0.086	0.005	0.143	0.699	0.006	11.579

TABLE 5.6 : Compilation time in hours when selecting variables on the whole dataset.

Learner	**c-index**		**iBrier**		**AUC score**		**selected**	**Time**
	Mean	sd	Mean	sd	Mean	sd		
SVSSA with $nIt = 5$	0.938	0.066	0.095	0.103	0.943	0.058	19	11.579
SVSSA with $nIt = 1$	**0.973**	0.059	**0.046**	0.041	**0.976**	0.051	28	8.774

TABLE 5.7 : Mean and standard deviation of performance measures obtained by 10-fold cross-validation on the training set for two configurations of SVSSA.

5.7.5 Cure models

Due to their convenience, we performed the performance analysis on Cox models. However, after a variable selection, the practitioner must build a predictive model with the estimation method that suits best their

problem. In the context of our study, where the event of interest includes relapse, disease progression and death, the hypothesis of the existence of a cured fraction in the population studied is obvious. Figure 5.11 also illustrates this point, although the follow-up time for some patients is too short to draw reliable conclusions. Cure models appear to be an obvious choice. We will not go into a deep analysis of TNBC

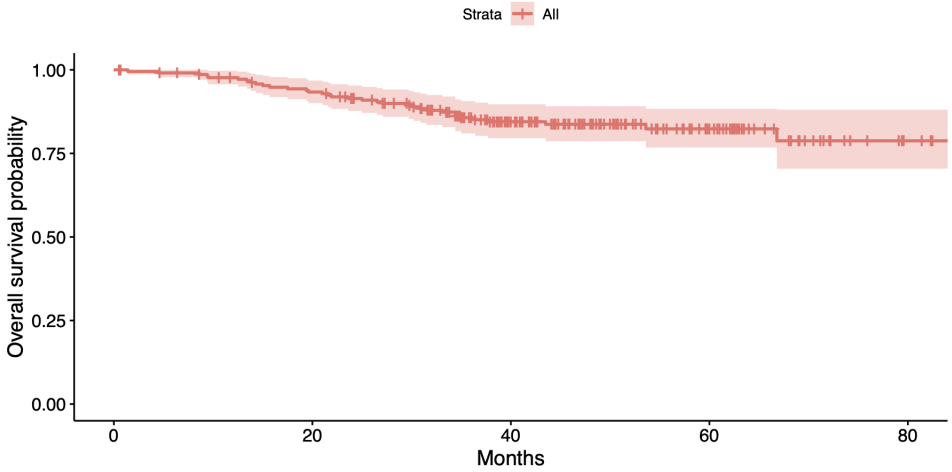


FIGURE 5.11 : Kaplan-Meier curve estimated with the whole dataset.

breast cancer by Cure models here, however, we offer in Figure 5.12 an overview of analyses that would be interesting to do next. In particular, it will be possible to compare the survival curves of individuals according to one or more variables, and so define the characteristics of the cured individuals, notably in terms of treatment received. As an example, we can see in Figure 5.12 that patients who have undergone a mastectomy will have a much higher probability of survival than patients who have not gone through this surgical intervention, all other things being equal.

5.8 Conclusion and perspectives

The SVSSA selection procedure aims at automating the different steps of variable selection in survival analysis. Based on a consensus between recognized methods, it is statistically consistent and offers good predictive performances, however its computation time limits its application for daily use. This shortcoming is obviously a prospect for improvement. One perspective would be to implement it in Python, but we would also have to think about other estimation methods available in this language. Moreover, the methods used are mostly built on the basis of a Cox model, alternatives based on Cure models could be considered to fit the context of clinical studies. Finally, we would like to specify that while our performance study was based on the dataset resulting from the second selection phase of step 4, we highly recommend to retrieve the list of the $P^\star$ most significant variables for an in-depth study of their relevance with TNBC specialists. These studies (selection with SVSSA and in-depth study of the selected variables) should also be performed on other omics datasets to evaluate the consistency of our algorithm.

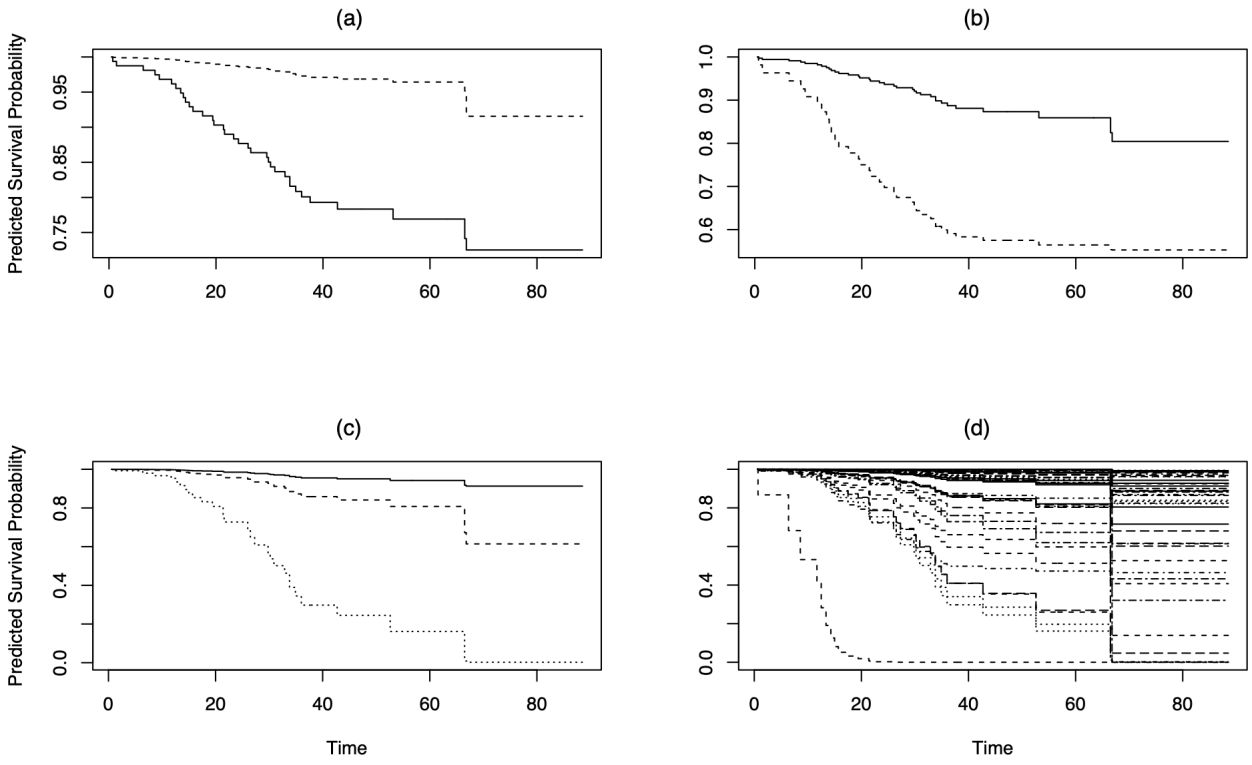


FIGURE 5.12 : (a) Survival curves of an individual who underwent a mastectomy (dashed line) and an individual who did not through this surgery (solid line) ; (b) Survival curves of an individual belonging to the iC4 subtype (dashed line) and an individual belonging to another subtype (solid line) ; (c) and (d) survival curves of different individuals for multivariate models.

Annexe A

Appendix: More details about the methods used in SVSSA

A.1 Preliminary step: Removing outliers

A.1.1 Residual analysis

For the next three methods, consider a Cox model characterized by a cumulative hazard function

$$H(t|X) = \int_0^t h_0(u)\exp(\beta X)\mathrm{du} = H_0(t)\exp(\beta X),$$

where β is the regression coefficients and H_0 is the cumulative baseline hazard function.

Martingale residuals[TGF90]. Introduced by Therneau *et al.*, the martingale residual of an individual i is given by

$$\hat{r}_{Mi} = \delta_i - \hat{H}_0(Y_i)\exp(\hat{\beta}X_i). \tag{1}$$

The residual $\hat{r}_{Mi}$ can be interpreted as the difference between the number of deaths observed over $[0, Y_i]$ by the individual i and the number expected given the model. Thus, a censored individual who, from the model's point of view, would have accumulated a significant risk over this period will be assigned a lower residual than if he had observed the event. In this sense, martingale residuals reveal individuals with different survival patterns from those with the same individual characteristics. However, taking its values in $]-\infty, 1]$, this specification has a marked skewness issue, therefore causing difficulty in anylizing its impact.

Deviance residuals [TGF90]. Deviance residuals have also been introduced by Therneau *et al.* in order to handle the skewness present in the martingale residuals distribution. More precisely, they propose a transformation of this measure in order to get as close as possible to a Gaussian distribution. To do so, the authors start from the definition of deviance in the general framework of linear regression models given by

$$D = 2\Big[\ell\ell(\beta^{(s)}) - \ell\ell(\hat{\beta})\Big],$$

where $\beta^{(s)}$ refers to the saturated model, which is a free model where each individual has its own vector of coefficients $\hat{\beta}_{.i}$; there are therefore no random error term. After an adaptation to survival models, and

in particular to Cox model, the authors define the deviance residual of an individual i by

$$\hat{r}_{Di} = \text{sign}(\hat{r}_{Mi})\sqrt{-2\left(\hat{r}_{Mi} + \delta_i \ln(\delta_i - \hat{r}_{Mi})\right)}, \tag{2}$$

where $\hat{r}_{Mi}$ corresponds to the martingale residual of this individual. We thus obtain a measure more centered around 0 with the same sign as the martingale residual. The residual $\hat{r}_{Di}$ can be interpreted as the difference in log-likelihood between an overfitted model and the retained model. We can deduce that individuals with extreme values on the residuals had a particular influence during the estimation of the fitted model ; they are potential outliers.

Censored quantile regression residuals [EHC14]. Eo *et al.* introduced the censored quantile regression residuals, adapting the Nardi and Schemper algorithm originally based on the Cox model to quantile regression for censored data. Consider $F_T(t) = \mathbb{P}(T \leqslant t)$ the distribution function of the survival time ; the τth quantile of T is defined as the inverse function

$$Q_T(\tau \,|\, X) = \inf\left\{ t : F_T(t \,|\, X) \geqslant \tau \right\} = \beta(\tau)\, X, \tag{3}$$

where $\beta(\tau)$ is the vector of regression coefficients at the quantile $\tau \in]0, 1[$. The authors define the deviance residual of an individual i by

$$\hat{r}_{Qi} = Y_i - Q(0.5 \,|\, X_i). \tag{4}$$

A.1.2 Predictive performance analysis

In survival analysis, the Concordance Index (C-index) is one of the most commonly used metrics to measure model performance. It quantifies the rank correlation between observed survival times and the risk levels defined by the fitted model. There are several algorithms for estimating the C-index, we will only present the most widespread one, that of Harrell *et al.* [HCP+82]. Specifically, the authors assess the model's ability to provide a reliable ranking of the survival times, by estimating the fraction of pairs of individuals correctly ordered out of all available comparable pairs. Let η_i denote the risk score that the individual i is assigned by the model. In the presence of censoring, we distinguish four cases:

- i and j have experienced the event: the pair (i, j) is correctly ranked if $\eta_i < \eta_j$ and $Y_i > Y_j$,
- i and j are censored: no information on the quality of the ranking can be provided,
- i is censored and j has experienced the event so that $Y_i > Y_j$: the pair (i, j) is correctly ranked if $\eta_i < \eta_j$,
- i is censored and j has experienced the event so that $Y_i < Y_j$: no information on the quality of the ranking can be provided.

The concordance index is then defined by

$$C_{\text{index}} = \frac{\sum_{i \neq j} \mathbb{1}_{\eta_i < \eta_j}\, \mathbb{1}_{Y_i > Y_j}\, \delta_j}{\sum_{i \neq j} \mathbb{1}_{Y_i > Y_j}\, \delta_j}. \tag{5}$$

It can be interpreted as the probability that an individual with a short survival time is considered by the model to be more at risk than an individual who has lived longer. We use this metric to determine the impact of each individual on the model predictive performances.

Dual Bootstraps Hypothesis Testing (DBHT) [PCV15a] Pinto *et al.* first introduced the Bootstraps Hypothesis Testing (BHT) method which assigns an anomaly score to each individual of the dataset, depending on whether they increase or decrease the model C-index [PCV15b]. Let $C_{\setminus i}$ be the C-index associated with the model built without the individual i, and C_{full} the one obtained with the complete dataset, the algorithm tests $\text{H}_0 : C_{\setminus i} \leqslant C_{\text{full}}$ on the basis of bootstrap samples from the dataset deprive of i. An outlier will therefore be characterized by a low p-value. However, the authors identified a limitation within this method. In particular, when the amount of available data is low, the bootstrap samples do not have enough information to lead to a relevant model, which could negatively impact the individuals' anomaly score. To overcome this constraint, the authors proposed the DBHT method which performs the comparison test on the basis of samples of the same size. Concretely for each individual i, the algorithm generates a set B_{poison} of bootstrap samples of the dataset that includes i in each bootstrap sample, and a set B_{antidote} that, conversely, is deprived of i. Next, the algorithm tests $\text{H}_0 : \mathbb{E}[C_{\text{antidote}}] > \mathbb{E}[C_{\text{poison}}]$ from the concordance histograms of the two sets. An outlier is again characterized by a low p-value.

A.1.3 Scoring

We follow the scoring procedure described by Carrasquinha *et al.* [CVV18], and based on the Rank Product test. In order to overcome the variability of the results offered by the various outlier detection methods, the authors propose to perform a consensus which takes the form of a scoring. For this, they perform the following rank product on each individual,

$$\text{RP}_i = \prod_{j=1}^{k} \text{rank}(Z_{ij}), \tag{6}$$

where k corresponds to the number of selected methods and Z_{ij} to the measure of outlyingness of individual i by method j. Note however that before applying the rank function, we harmonize the measures so that a low value of Z_{ij} corresponds to a high level of outlyingness for the method j. Thus, individuals with low ranks, and by continuity a low RP_i, would a priori be more outlier. The algorithm then performs a significance test of this ranking based on H_0 : the ranking is random. The anomaly score is then represented by this test's p-value, the lower it will be, the more the individual will be considered as an outlier. In other words, if an individual is systematically defined as an outlier by the selected methods, it is likely that this classification is correct.

A.2 Step 1: Marginal selection

A.2.1 Correlation with survival time

To avoid any bias induced by censoring, the correlation analysis with survival time is only based on individuals who observed the event, *i.e.* $(\mathbf{X}^{(d)}, \mathbf{Y}^{(d)})$.

Spearman correlation [Spe04] The Spearman rank correlation coefficient is a nonparametric measurement of dependence based on the rank statistic. Let X_j be a quantitative predictor and T the actual survival time variable, for which we have the sample $(\mathbf{X}_j^{(d)}, \mathbf{Y}^{(d)})$. The Spearman correlation coefficient between $\mathbf{X}_j^{(d)}$ and $\mathbf{Y}^{(d)}$ is given by

$$\rho_j = \frac{S\big(\text{rank}(\mathbf{X}_j^{(d)}),\, \text{rank}(\mathbf{Y}^{(d)})\big)}{\sqrt{S\big(\text{rank}(\mathbf{X}_j^{(d)})\big)\, S\big(\text{rank}(\mathbf{Y}^{(d)})\big)}}, \tag{7}$$

where $S(.)$ and $S(.,.)$ correspond respectively to the empirical variance and covariance functions. RS_j indicates the degree of monotonic dependence between the two variables ; the closer it is to 1 in absolute value, the higher the dependence. However, we are not interested in the correlation coefficient itself but in its significance. We then proceed with a bilateral test verifying $\text{H}_0 : RS_j = 0$. A low p-value indicates the existence of a potential monotonic dependence between X_j and T. This measure constitutes our significance score for this method.

A.2.2 Correlation with survival status

Here we analyze the correlation with survival status δ from the sample $(\delta, \mathbf{X})$.

Wilcoxon–Mann–Whitney test [Wil45, MW47] The W-M-W test is a nonparametric statistical hypothesis test based on the rank statistic. It verifies the existence of significant differences between two groups of individuals on the basis of a quantitative variable. Consider a quantitative predictor X_j for which we have independent and identically distributed observations, distributed within two samples $\mathbf{X}_j^{(d)}$ and $\mathbf{X}_j^{(c)}$. Let F_d and F_c be the distribution functions corresponding to the two samples. The bilateral W-M-W test verifies $\text{H}_0 : F_d(t) = F_c(t)\ \forall t$. A low p-value implies a significant difference between the two groups, and therefore the existence of a potential dependence between X_j and δ. This measure constitutes our significance score for this method.

A.2.3 Correlation with the survival pair

The last method of this step studies the correlation between each variable and the pair (Y, δ). To this end, we use two measures of model performance, the Uno's concordance index (c-index) and the integrated Brier score (iBrier) ; while the c-index only assesses the model discrimination power, the iBrier evaluates also its accuracy. By using these two measures, we want to exploit the idea that a significant variable will result in a well-performing model.

Uno's concordance index [UCP+11]. Uno *et al.* shown that the Harrell index presented in A.1.2 is biased when the number of censored data is high, and proposed an adaptation that overcomes this shortcoming. To assess the relationship between a variable j and the survival pair, we compute the c-index of the simple cox model (X_j is the only predictor) by 10-fold cross-validation. This measure constitutes our significance score for this method.

Integrated Brier score [GSSS99]. The Brier score was initially proposed for uncensored data. In this setting, it represents the mean squared deviation between the observed survival status and the survival probability estimated by the model. Graf *et al.* have later adapted it to survival problems with right censoring by weighting the squared deviations by the inverse probability of censoring. Let $G(t) = \mathbb{P}(C > t)$ be the survival function of the censoring times, *i.e* the probability of not being censored until t. The Brier score at a given time t is defined by

$$BS(t) = \frac{1}{n} \sum_{i=1,\dots,n} \begin{cases} \frac{\left(0 - \hat{S}(t|X_i)\right)^2}{\hat{G}(t_i)} & \text{if } t_i \leqslant t,\ \delta_i = 1 \\ \frac{\left(1 - \hat{S}(t|X_i)\right)^2}{\hat{G}(t)} & \text{if } t_i > t \\ 0 & \text{if } t_i = t,\ \delta_i = 0, \end{cases} \tag{8}$$

with $\hat{G}(t)$ the Kaplan-Meier estimator of the censoring distribution. The integrated form provides an overall measure and is given by

$$IBS = \frac{1}{t_n} \int_0^{t_n} BS(t)\ \mathrm{dt}. \tag{9}$$

A.2.4 Scoring

The global scoring is based on the Rank Product test procedure described in Section A.1.3. After harmonizing the scores, so that a low score refers to a high significance of the variable, the scoring function returns the p-value of associated with its ranking. For each omic group g, we select the $P_g^{(1)}$ variables with the lowest p-values.

A.3 Step 2: Selection of correlated groups

A.3.1 Hierarchical clustering

We aim to group together the correlated variables within the estimation procedures, while limiting the size of the clusters to spare the borrowed alrogithms. We chose to perform a hierarchical clustering based on the proximity distance provided by unsupervised Random Forest.

Unsupervised Random Forest [BC03] Breiman and Cutler proposed an unsupervised version of the well-known Random Forest algorithm. The approach consists in returning to the standard classification case by creating a synthetic target variable. Let D_O be the initial unlabeled dataset, the algorithm constitutes a synthetic dataset D_S, of the same size as D_O, by random sampling from the product of empirical marginal distributions of the predictors. The target variable is then created by assigning class 1

to the original data, and class 2 to the synthetic data. Thus, class 2 follows a distribution of independent random variables and does not have the correlation structure of the original data. A Random Forest classification model is then constructed based on the new data sample. The algorithm aims to obtain the finest possible predictor, *i.e.* the one that best separates the noise from the real data. Therefore, the trees must fit the correlation structure present in the original data, based on the dependent variables. In this sense, a model with a low classification error rate illustrates the presence of correlation within the observations ; the most correlated of them should end up in the same leaf nodes. The unsupervised Random Forest then provides a proximity matrix P which offers an estimate of the distance between individuals according to the proportion of times they are found in same leaves. Several studies have shown the interest of P and its advantages for clustering [CS06, SSB+05, AHT+03, KWB18].

In our algorithm, we use this method to obtain the proximity matrix P_g of the predictors of each initial group g. We then consider the transpose of the dataset restricted to the predictors of g, so that these represent the individuals to be classified in unsupervised trees.

Ward hierarchical clustering [WJ63] Hierarchical ascending clustering (HAC) is an iterative grouping method based on a dissimilarity measure. We will use Ward's HAC which is the most widespread. The algorithm is initialized with as many clusters as there are individuals, and merges at each step the two closest clusters in order to minimize the intra cluster variance, which is defined using a distance matrix between individuals. Let P_g be the proximity matrix of the predictors of group g. Let $\bar{P}_g = 1 - P_g$ be the distance matrix of the said predictors (which are our individuals here). The CAH optimization problem is given by

$$\underset{A,\, B \in C_{\text{actual}}}{\arg\min} \frac{1}{n_{A\cup B}} \sum_{x,\, y \in A\cup B} \bar{P}_g(x,y) - \frac{1}{n_A} \sum_{x,\, y \in A} \bar{P}_g(x,y) - \frac{1}{n_B} \sum_{x,\, y \in B} \bar{P}_g(x,y), \tag{10}$$

where C_{actual} is the set of current clusters, n_X the number of elements in cluster X, and $\bar{P}_g(x,y)$ the distance between x and y. The algorithm stops when there is only one cluster left. Once the maximal tree is obtained, we split it in order to recover the desired partition in m_g clusters.

A.3.2 Estimation with survival time

As before, the estimation with the survival time will be based on the sample $(\mathbf{Y}^{(d)}, \mathbf{X}^{(d)})$ to avoid the bias induced by censoring. We use here our Bayesian approach of the PGGM under the group-sparse setting.

Bayesian PGGM (see Chapter 3) For each set of four predictor subgroups, we apply the Bayesian PGGM algorithm with a default parameterization inducing sparsity, which is: $a = 100$, $b = 1$, type = group-sparse, and shrinkage = adaptive. We have chosen not to launch a parameter refinement process in order not to increase the computation time. However, we leave to the practitioner the possibility to fill in the parameters that suit him. The resulting model returns the list of subgroups selected from the four candidates. The latter then gain a point of significance.

A.3.3 Estimation with survival status

We continue with a classification-type variable selection method based on the sample $(\delta, \mathbf{X})$. The method borrowed here is the standard logistic group Lasso.

Logistic Group Lasso[MVDGB08] The group Lasso of Yuan and Lin is an extension of the Lasso that performs variable selection at the group level in linear regression models. Meier *et al.* proposed an adaptation of this method for logistic regression. Let $X = (X_1^t, ..., X_m^t)^t \in \mathbb{R}^p$ the vector of predictors divided into m groups, let p_g be the size of X_g the vector of group g. The logistic model is defined by

$$\mathbb{P}(\delta = 1 | X = X_i) = \frac{1}{1 + \exp\big(-\beta_0 - \sum_{g=1}^m \beta_g^t X_{gi}\big)}, \tag{11}$$

where $\beta = (\beta_0, \beta_1^t, ..., \beta_m^t)^t \in \mathbb{R}^{p+1}$ is the vector of regression coefficients. The group Lasso estimation of β is based on the following optimization problem

$$\hat{\beta}_\lambda = \arg\min_\beta -\ell\ell(\beta) + \lambda \sum_{g=1}^m s(p_g)|\beta_g|_2, \tag{12}$$

where $\ell(\beta)$ is the log-likelihood of the logistic model. Regarding to the penalty function, $s(p_g)$ is a weight function which allows to rescale the penalty with respect to the size of the group. The default parameter is $s(p_g) = \sqrt{p_g}$, but this function can be modulated, for example it can be set to 0 for a group that we do not want to penalize. In our context, we consider four subgroups on which we apply the logistic group Lasso algorithm with a default parameterization inducing sparsity, i.e. $\lambda = 10^{-2}$. Here again, we do not run a cross-validation procedure to adjust the parameter in order not to increase the computation time. The practitioner can of course change the default parameter. Just like for the Bayesian PGGM, the created model returns the list of the selected subgroups, those ones get a significance point.

A.3.4 Estimation with the survival pair

For this substep, we will use two methods specific to survival analysis that incorporate group structures. The support sample is $(\mathbf{Y}, \delta, \mathbf{X})$.

IPF-Lasso [BDBJF17] The IPF-Lasso method has been proposed by Boulesteix *et al.* as an improvement of Lasso in the context of multi-omics survival analysis. The authors explain that the use of several types of high-dimensional data requires the implementation of more adaptive selection methods. Depending on the case, clinical data, which is both very small in size and very significant for prognosis, should not be penalized in the same way as very large omics data such as copy numbers. Therefore, their method allows to assign different penalty factors to groups of variables. The optimization problem follows that of the Cox Lasso, incorporating the group structure and different penalty terms

$$\hat{\beta} = \arg\min -\ell\ell_p(\beta) + \sum_{g=1}^m \lambda_g |\beta_g|_1, \tag{13}$$

where $\ell\ell_p(\beta)$ is the partial likelihood of the Cox model. This method adapts perfectly to linear and logistic regressions, by modifying the likelihood term. The resulting model is a selection at variable level but conditional on the group of membership. In our context of four subgroups, we build the model using the cross-validation algorithm proposed by the authors. Once the model is obtained, we consider a subgroup as significant if none of its variables has been penalized.

Block Forest [HW19] The Block Forest method proposed by Hornung and Wright is a variant from the well-known Random Forest by Breiman [Bre01]. Inspired by the multi-omics data problem, this method considers variables group structure at the level of the split point selection. Recall that a random forest model is a set of decision trees built from bootstrap samples of the training data. Each decision tree is a series of binary tests that aim to divide the individuals according to a criterion of homogeneity in regard to the output. The point of interest here is the construction of these binary tests, or more precisely the split point selection procedure. In the Random Forest algorithm, this is done by a random selection of *nvar* predictors on which the algorithm selects the division that optimizes the considered split criterion. Hornung and Wright pointed out that treating each variable uniformly, when in a group structuring, and especially in the multi-omics framework, large groups will be overrepresented in the model regardless of their relevance as predictors. To overcome this shortcoming, the authors propose a weight-based splitting procedure. First, in order to give each group the possibility to be represented within the split point selection, the algorithm draws each group with a probability of $\frac{1}{2}$, *i.e.* all groups are selected with a probability $\left(\frac{1}{2}\right)^m$, and the draw is rerun if no group is selected. Then, for each admitted group g of size p_m, the algorithm randomly selects $\sqrt{p_g}$ predictors to be candidates for the split, so $nvar = \sum_{g \text{ admitted}} \sqrt{p_g}$. Finally, the split criterion is weighted by the weights w_g associated to each group in order to privilege variables coming from groups supposed to be more influential. The weights $(w_1, ..., w_m)$ are defined by the algorithm through an optimization process. The optimal weights associated with a model provide an indication of the relative importance of the different groups for prediction.
This method adapts to all variants of the Random Forest, including Ishwaran's random survival Forest *et al.* [IKBL08] that we are interested in here. Since we wish to achieve a group selection, once our model is built from the four predictor subgroups, we consider a subgroup significant if its optimized weight is greater than a threshold set at 0.75 by default, but configurable by the practitioner.

A.3.5 Scoring

The global scoring consists simply in counting the number of times that a given subgroup has been selected by the different methods. We then select the 10% most significant ones. Since they have different sizes, this is not equivalent to the 10% most significant variables.

A.4 Step 3: Selection of decorrelated variables

A.5 Estimation with survival time

As before, the estimation with the survival time will be based on the sample $(\mathbf{Y}^{(d)}, \mathbf{X}^{(d)})$ to avoid the bias induced by censoring. We use here our structuring approach of the PGGM.

Structural PGGM (see Chapter 2) For each subset of data, we apply the GenGm algorithm with a default parameterization inducing sparsity, which is: $\lambda = 0.001$, $\mu = 0.05$, $\eta = 0.001$, and $\beta = 1.25$. The matrix $L = C^{-1}$ associated with the structuring penalty is constructed from the subgroups obtained by clustering in step 2. For each pair of inputs (i, j) we have

$$C_{ij} = \begin{cases} 1 & \text{if } i = j, \\ \frac{1}{10} & \text{if } i \text{ and } j \text{ are in the same subgroup,} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

The resulting model returns the list of inputs selected from the *nvar* candidates, these gain a point of significance.

A.5.1 Estimation with survival status

In the continuity of the group Lasso of step 2, we now borrow the well-known Lasso method on the sample $(\delta, \mathbf{X})$.

Logistic Lasso [Lok99, SK03] The Lasso method offers a sparse estimation of regression models, through a penalty ℓ_1 on the vector of coefficients. In the logistic framework, it is based on the following optimization problem

$$\hat{\beta}_\lambda = \arg\min_\beta -\ell\ell(\beta) + \lambda \sum_{j=1}^{p^{(o)}} |\beta_j|, \tag{15}$$

where $\ell\ell(\beta)$ is the log-likelihood of the logistic model, and λ to the regularization parameter. In our procedure, we set this parameter by default to 10^{-2}. At each iteration the model obtained returns the list of selected inputs, these gain one point of significance.

A.5.2 Estimation with the survival pair

For this substep, we will use two methods designed for variable selection, a boosting and a bayesian approaches. The support sample is $(\mathbf{Y}, \delta, \mathbf{X})$.

Gradient boosting with component-wise for Cox model [Bue06, DB16] Boosting methods are based on an iterative estimation of model parameters, in order to gradually reduce its loss function. More precisely, they aim to build a robust model from a set of *weak* models. There are several variants categorized according to the *weak* model estimation procedures and the final model construction rules. We

are interested here in the gradient-based method within Cox model framework. In this context, we seek to estimate the regression coefficients β by minimizing the negative partial log-likelihood which constitutes our loss function. The algorithm is initialized with $\beta = 0_{p,1}$, then for each iteration, it computes the negative gradient vector of the loss function with respect to $\eta(X, \beta) = X^t\,\beta$. For each individual i, it is given by

$$\begin{aligned} U_i &= \frac{\partial}{\partial \eta(\mathbf{X}_{\cdot i}, \beta)}\, \ell\ell_p(\beta))\Big|_{\eta(X,\beta)=\eta(X,\hat{\beta})} \\ &= \delta_i - \sum_{l \in R(y_i)} \frac{\exp(\mathbf{X}_{\cdot l}^t\, \hat{\beta})}{\sum_{k \in R(y_l)} \exp(\mathbf{X}_{\cdot k}^t\, \hat{\beta})}. \end{aligned} \tag{16}$$

In order to maximize the partial likelihood (*i.e.* minimize the loss function) in the direction most correlated with the gradient, the algorithm estimates a linear model fitting the gradient U to each input X_j, that is

$$\hat{b}_j = \arg\min_{\beta_j} \frac{1}{n} |U - \mathbf{X}_{j\cdot}\, \beta_j|_2^2. \tag{17}$$

The *weak* model associated with the current iteration is then denoted by the coefficient $\hat{b}_{j^\star}$ for which $j^\star = \arg\min_j \ \frac{1}{n}|U - \mathbf{X}_{j\cdot}\, \hat{b}_j|_2^2$. Finally, the algorithm updates the estimator $\hat{\beta}_{j^\star} = \hat{\beta}_{j^\star} + \nu\, \hat{b}_{j^\star}$ where ν controls the learning rate. This phase of estimating *weak* models is iterated It_{max} times. This number can be used as a regularization parameter to perform variable selection. In our procedure, we use the default parameters defined in `R` which are: $\nu = 0.1$ and $It_{max} = 100$.

Spike-and-Slab Lasso Cox [TSZY17] Proposed by Tang *et al.*, the Spike-and-Slab Lasso Cox is a Bayesian variables selection method offering an adaptive shrinkage on Cox model regression coefficients. This hierarchical model is based on a spike-and-slab double-exponential prior (*i.e* a Laplace distribution), and is defined as follows

$$\begin{cases} \beta_j | \pi_j, s_0, s_1 & \sim & (1-\pi_j)\, \mathcal{L}(\beta_j | 0, s_0) + \pi_j\, \mathcal{L}(\beta_j | 0, s_1) \\ \pi_j | \theta & \sim & \mathcal{B}(\theta) \\ \theta & \sim & \mathcal{U}(0,1) \end{cases} \tag{18}$$

where π_j is the prior spike probability, and s_0 and s_1 are predefined scale parameters such as $0 < s_0 < s_1$. s_0 is the spike scale parameter chosen small to induce strong shrinkage on β_j, and s_1 is the slab scale parameter chosen large to instead induce weak or no shrinkage on the coefficient. The sparse model is then estimated by searching for the posterior modes of the parameters, in particular by maximizing the log joint posterior density which verifies

$$\ln p(\beta, \pi, \theta \,|\, \mathbf{Y}, \delta, \mathbf{X}) \propto \ell\ell_p(\beta) - \sum_{j=1}^{p} \frac{1}{S_j} |\beta_j| + \sum_{j=1}^{p} (1-\pi_j)\, \ln(1-\theta) + \pi_j \ln(\theta), \tag{19}$$

where $\ell\ell_p(\beta)$ is the Cox model partial log-likelihood, $S_j = (1-\pi_j)s_0 + \pi_j s_1$, and the term $\sum_{j=1}^{p} \frac{1}{S_j} |\beta_j|$ can be likened to a Lasso-type penalty. In our procedure we set by default $s_0 = 0.2$ and $s_1 = 0.5$.

A.5.3 Scoring

The scoring procedure for this step is based on the *minimum redundancy maximum relevance* approach of De Jay *et al.* [DJPCO+13]. It consists in an iterative update of candidate variable importance scores according to their degree of association with the previously selected ones. Let's start by assigning each variable j a basic score of significance, which corresponds to the number of times it was designated significant during the estimation phase. Let us denote this score $R_j^{(0)}$ and let S be the set of selected variables. The algorithm initializes S with the variable having the highest base score, then updates the individual scores by penalizing them by the degree of correlation with the selected variables. During the following iterations, S is completed by adding the variable maximizing the new score ; in other words, the one that verifies

$$j^\star = \arg\max_{j \notin S} R_j^{(0)} - \frac{1}{\mathrm{card}(S)} \sum_{l \in S} \rho_S(\mathbf{X}_{j\cdot}, \mathbf{X}_{\cdot l}), \tag{20}$$

where $\rho_S(\mathbf{X}_{j\cdot}, \mathbf{X}_{\cdot l})$ denotes the Spearman coefficient between inputs j and l. The algorithm stops when the size of S reaches the desired number of selected feautures.

A.6 Step 4: Final selection

A.6.1 Estimation with the survival pair

For this estimation phase, we use four variable selection methods, a group structure approach, an approach favoring clinical variables, a nonlinear approach by decision trees, and a boosting approach. The support sample is $(\mathbf{Y}, \delta, \mathbf{X})$.

Sparse-Group Lasso [SFHT13] This method proposed by Simon *et al.* is an extension of the Lasso which performs a variable selection both on group and within group level. It is based on the combination of two forms of penalties ; a ℓ_2-norm penalty to induce groupwise sparsity (like the Group-Lasso), and a ℓ_1-norm penalty to induce intra-group sparsity within non-zero groups (like a Lasso per group). In the Cox model framework, the estimation is based on the following optimization problem

$$\hat{\beta}_\lambda = \arg\min_\beta -\ell\ell(\beta) + (1-\alpha)\,\lambda \sum_{g=1}^{m} \sqrt{p_g}\,|\beta_g|_2 + \alpha\,\lambda\,|\beta|_1. \tag{21}$$

In our procedure we tune the penalty parameter λ by an 10-fold cross-validation, and set by default $\alpha = 0.95$ to encourage grouping as recommended by the authors.

priority-Lasso [KJH+18] Klau *et al.* emphasize the interest for practitioners in having high-dimensional estimation methods that incorporate variable group structure and exploit a prior knowledge on their practical usability. Thus, in the context of omics data, for approximately the same degree of precision, practitioners will choose a model built from variables already included in routine diagnostics such as clinical data, than variables with a high acquisition cost. To handle these two points, the authors proposed the priority-Lasso method. This is a hierarchical regression method that takes group structures into account, which are ordered according to a prior knowledge of their priority. The idea is to update

the estimator an iterative way by building, at each step, a predictor explaining the current estimator residuals with the next priority group. Thus, groups with low priority variables will only enter the model if they explain a variability of the outcome that is not explainable by groups with higher priority. Let us place ourselves in the Cox model framework, and let $\pi = (\pi_1, ..., \pi_m)$ be the permutation of the groups according to their order of priority (*i.e.* π_1 is the group with the highest priority). The first step of the estimation procedure consists in extracting the information available in the highest priority group. To do this, the algorithm builds a Lasso predictor on the output with variables from π_1, by solving

$$\hat{\beta}^{(\pi_1)} = \underset{\beta^{(\pi_1)}}{\arg\min} -\ell\ell_p(\beta^{(\pi_1)}|\mathbf{Y}, \delta, \mathbf{X}^{(\pi_1)}) + \lambda^{(\pi_1)} |\beta^{(\pi_1)}|_1, \tag{22}$$

where $\ell\ell_p(\beta^{(\pi_1)}|\mathbf{Y}, \delta, \mathbf{X}^{(\pi_1)})$ corresponds to the log-likelihood of the Cox model restricted to variables from π_1. Let $h^{(\pi_1)}$ be the predictor associated with $\hat{\beta}^{(\pi_1)}$ and $r^{(\pi_1)}$ its observed residuals. The following steps seek to gradually explain the remaining information on the output variability using increasingly lower priority groups. At each step g is then constructed a Lasso predictor on the current estimator residuals according to the next priority group, by solving

$$\hat{\beta}^{(\pi_g)} = \underset{\beta^{(\pi_g)}}{\arg\min} -\ell\ell_p(\beta^{(\pi_g)}|r^{(\pi_{g-1})}, \mathbf{X}^{(\pi_g)}) + \lambda^{(\pi_g)} |\beta^{(\pi_g)}|_1. \tag{23}$$

The new estimator $h^{(\pi_g)}$ is associated with the coefficient vector $(\hat{\beta}^{(\pi_1)}, ..., \hat{\beta}^{(\pi_g)})$. The final estimator is obtained when the m groups have been considered. Thus constructed, a non-priority group will only be integrated into the model if it offers non-redundant information with the previous groups. However, the authors raised a shortcoming about this procedure. Since the residuals are computed on the data that were used to train the predictors, the model accuracy may be overestimated. To overcome this, they propose to estimate each predictor by cross-validation. We follow this recommendation in our procedure and also determine the penalty parameters $\lambda^{(\pi_g)}$ by cross-validation.

Random Survival Forest [IKBL08] In the paragraph on the Block Forest method, we have slightly discussed the construction of random survival Forest. This method, proposed by Ishwaran *et al.*, is an adaptation of Breiman's Random Forest [Bre01] to survival data. Like the reference method, it is based on the construction of B survival trees using bootstrap samples from the original data. Each tree is a set of binary tests dividing the population into nodes more and more homogeneous, until they reach terminal nodes. At each node split, the algorithm randomly draws *nvar* candidate variables, then selects among them the split that maximizes survival difference between daughter nodes. The growth of a tree stops when it reaches a terminal node containing a homogeneous population, or a minimum of $d_0 > 0$ deaths. A survival Forest aims to estimate an ensemble cumulative hazard function from each individual tree predictions. More specifically, each terminal node h is associated with the following cumulative hazard function

$$\hat{H}_h(t) = \sum_{\mathbf{Y}_{\cdot i}^{(h)} < t} \frac{\delta_i^{(h)}}{R(\mathbf{Y}_{\cdot i}^{(h)})}, \tag{24}$$

where $(\mathbf{Y}^{(h)}, \delta^{(h)})$ is the sample of observations in h and $R(\mathbf{Y}_{.i}^{(h)})$ the number of individuals, within h, still at risk at time $\mathbf{Y}_{.i}^{(h)}$. Thus, to each individual $i \in h$ is associated $\hat{H}(t|\mathbf{X}_{.i}) = \hat{H}_h(t)$. The ensemble cumulative hazard function of an individual i is then defined by

$$\hat{H}_{RSF}(t|\mathbf{X}_{.i}) = \frac{1}{B}\sum_{b=1}^{B} \hat{H}_b(t|\mathbf{X}_{.i}). \tag{25}$$

In our procedure, we use the default parameters *nvar*, B and d_0 provided in the `R` algorithm, and for variable selection purposes, we retrieve the variable importance matrix from the model. The $P^\star$ most significant variables are considered significant for this substep.

Likelihood-based boosting for Cox model [TB06, BASB09, DB16] Tutz and Binder proposed a boosting method suitable for any generalized linear model. The estimation procedure is based on a e componentwise approach, where at each step one coefficient of the estimator is updated by a *weak* model. In the Cox model framework, the loss function to be minimized is defined by the ℓ_2-norm penalized negative partial log-likelihood, in which the results of the previous boosting steps are incorporated as an offset. Thus at step k, the *weak* model of a variable j results from the following loss function

$$\ell\ell_{\text{pen}}(\beta_j|\hat{\beta}) = -\sum_{i=1}^{n} \delta_i \left[\hat{\eta}_i + \mathbf{X}_{.i}^t \beta_j - \ln\Big(\sum_{l\in R(y_i)} \exp\left\{\hat{\eta}_l + \mathbf{X}_{.l}^t \beta_j\right\}\Big)\right] + \frac{\lambda}{2}\beta_j^2, \tag{26}$$

where $\hat{\beta}$ is the estimator updated at step $k-1$, and $\hat{\eta} = \mathbf{X}^t\hat{\beta}$ the offset term from this step. The algorithm is initialized with $\hat{\beta} = 0_{p,1}$ and $\hat{\eta} = 0_{n,1}$. Then, for each iteration, it first computes $\hat{b}_j$ the first-order approximation around 0 of (26) for each variable j. This is given by

$$\hat{b}_j = \frac{\partial}{\partial\beta_j}\ell\ell_{\text{pen}}(0|\hat{\beta})\left(\frac{\partial^2}{\partial\beta_j^2}\ell\ell_{\text{pen}}(0|\hat{\beta})\right)^{-1}. \tag{27}$$

The *weak* model associated with the current iteration is then designated by the coefficient $\hat{b}_{j^\star}$ for which $j^\star = \arg\min_j \ell\ell_{\text{pen}}(\beta_j|\hat{\beta})$. Finally, the algorithm updates the estimator $\hat{\beta}_{j^\star} = \hat{\beta}_{j^\star} + \hat{b}_{j^\star}$. This phase of estimating *weak* models is iterated It_{max} times. This number is used as a regularization parameter to perform variable selection. In our procedure, we use the default parameter $It_{max} = 2P^\star$.

A.6.2 Selection

As previously mentioned, the selection phase of this final step is divided into two parts. The first part is based on a scoring which consists in counting the number of times each variable has been selected during the estimation phase. The $P^\star$ variables with the highest score are selected. The second phase of selection consists in searching for the best performing combination, among the most significant variables. First, we perform a bottom-up selection procedure ; at each step we add a variable to the model if it reduces the iBrier. Then we perform a top-down selection on the obtained model ; at each step we remove a variable from the model if this removal reduces the iBrier. The iBrier scores are obtained by cross-validation on the training data. Moreover, even if at the end of the procedure we obtain a combination of performing

variables, we nevertheless recommend that practitioners also study the list of $P^\star$ variables selected by scoring.

Appendix: Variables selected by SVSSA

B.1 100 most significant variables

Matrices	Variables
Clinical	$LN\_positive$, $Surgery\_BCS$, $iCluster\_Subtype\_iC4$, $Surgery\_Mastectomy$, $Chemotherapy\_FAUX$, $Size\_cm$, $Ascat\_Ploidy$, $Chemotherapy\_VRAI$, $CNA\_Subtype\_Chr13q34\_amp$, $Chemotherapy\_other\_regimen\_VRAI$, $iCluster\_Subtype\_iC1$, $iCluster\_Subtype\_iC9$, $Ki67$, N_2, $Chemotherapy\_other\_regimen\_FAUX$, $mRNA\_Subtype\_IM$, $iCluster\_Subtype\_iC7$
RNAseq	$RSL24D1P6$, $ATP7B.x$, $KCNS2.x$, $RP11.93K22.12$, $RP11.82L18.2$, $NCAM2.x$, $RP11.697N18.2$, $RP11.641D5.1$, $PLA2G16.x$, $NOX4.x$, $INHA$, $AC009996.1$, $CNPY3$, $ZNF625$, $BHLHE22.x$, $RSL24D1$, $RP11.384C4.7$, $AL117352.1$, $RP13.204A15.3$, $RP11.38L15.2$, $HIST1H3G$, $PPP1R15B$, $NIPAL2$, $CTD.2085F10.1$, $UNC50.x$, $HSPB11.x$
CNA	$S.tag298342$, $S.tag020770$, $S.tag280437$, $S.tag195468$, $S.tag280382$, $S.tag292697$, $S.tag227455$, $S.tag041737$, $S.tag306109$, $S.tag000842$, $S.tag264435$, $S.tag026817$, $S.tag250687$, $S.tag223243$, $S.tag286210$, $S.tag141796$, $S.tag282804$, $S.tag069019$, $S.tag103995$, $S.tag091076$, $S.tag123168$, $S.tag080746$, $S.tag277180$, $S.tag014996$, $S.tag029426$, $S.tag191575$, $S.tag030768$, $S.tag317958$, $S.tag131246$, $S.tag161315$, $S.tag030729$, $S.tag049559$, $S.tag016873$, $S.tag267337$, $S.tag213248$, $S.tag236926$, $S.tag083564$, $S.tag209362$, $S.tag011299$, $S.tag104795$, $S.tag032493$, $S.tag011156$, $S.tag214732$, $S.tag199561$, $S.tag196229$, $S.tag121446$, $S.tag231311$, $S.tag013472$, $S.tag207164$, $S.tag225078$, $S.tag144768$, $S.tag101438$, $S.tag230101$, $S.tag083900$, $S.tag097908$, $S.tag167905$, $S.tag106361$

B.2 Variables selected after the forward-backward procedure

Matrices	Variables
Clinical	$LN\_positive$, $iCluster\_Subtype\_iC4$, $Surgery\_Mastectomy$
RNAseq	$RSL24D1P6$, $RP11.82L18.2$, $RP11.641D5.1$, $INHA$, $ZNF625$, $NIPAL2$
CNA	$S.tag298342$, $S.tag020770$, $S.tag280437$, $S.tag195468$, $S.tag227455$, $S.tag041737$, $S.tag026817$, $S.tag080746$ $S.tag030768$, $S.tag317958$

BIBLIOGRAPHY

[AG07] G. Andrew and J. Gao. Scalable training of L_1-regularized log-linear models. *Proc. 24th Inte. Conf. Mach. Learning.*, pages 33–40, 2007.

[AHT+03] Elena Allen, Steve Horvath, Frances Tong, Peter Kraft, Elizabeth Spiteri, Arthur D Riggs, and York Marahrens. High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proceedings of the National Academy of Sciences*, 100(17):9940–9945, 2003.

[APN+03] Anthony Antoniou, Paul DP Pharoah, Steven Narod, Harvey A Risch, Jorunn E Eyfjord, John L Hopper, Niklas Loman, Håkan Olsson, O Johannsson, Åke Borg, et al. Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics*, 72(5):1117–1130, 2003.

[BASB09] Harald Binder, Arthur Allignol, Martin Schumacher, and Jan Beyersmann. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, 25(7):890–896, 2009.

[BBV04] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[BC03] Leo Breiman and Adele Cutler. *Random Forests Manual V4.0*. Technical report, UC Berkeley, 2003.

[BDBJF17] Anne-Laure Boulesteix, Riccardo De Bin, Xiaoyu Jiang, and Mathias Fuchs. Ipf-lasso: integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and mathematical methods in medicine*, 2017, 2017.

[BDDW08] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[BEGD08] O. Banerjee, L. El Ghaoui, and A. D'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.

[BMA+20] R. Bai, G. E. Moran, J. L. Antonelli, Y. Chen, and M. R. Boland. Spike-and-slab group lassos for grouped regression and sparse generalized additive models. *J. Am. Stat. Assoc.*, pages 1–14, 2020.

[Boa49] John W Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53, 1949.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[Bro22] Jason Brownlee. Data preparation for machine learning, 2022.

[Bue06] Peter Buehlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583, 2006.

[BVF98] P. J. Brown, M. Vannucci, and T. Fearn. Multivariate Bayesian variable selection and prediction. *J. R. Statist. Soc. B.*, 60(3):627–641, 1998.

[Chr87] Erik Christensen. Multivariate survival analysis using cox's regression model. *Hepatology*, 7(6):1346–1358, 1987.

[CLL11] T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.*, 106(494):594–607, 2011.

[CMHR17] J. Chiquet, T. Mary-Huard, and S. Robin. Structured regularization for conditional Gaussian graphical models. *Stat. Comput.*, 27(3):789–804, 2017.

[Cox75] David R Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[CS06] Adele Cutler and John R Stevens. [23] random forests for microarrays. *Methods in enzymology*, 411:422–432, 2006.

[CVV18] Eunice Carrasquinha, André Veríssimo, and Susana Vinga. Consensus outlier detection in survival analysis using the rank product test. *bioRxiv*, page 421917, 2018.

[DB16] Riccardo De Bin. Boosting in cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the r-packages coxboost and mboost. *Computational Statistics*, 31(2):513–531, 2016.

[dC] Institut National du Cancer. `https://www.e-cancer.fr/Professionnels-de-sante/Les-chiffres-du-cancer-en-France/Epidemiologie-des-cancers/Les-cancers-les-plus-frequents/Cancer-du-sein`. Accessed: 16-03-2022.

[dCdl] Institut de Cancérologie de l'Ouest. `https://www.institut-cancerologie-ouest.com/cancer-du-sein`. Accessed: 16-03-2022.

[Dem72] Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

[DJPCO+13] Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, and Benjamin Haibe-Kains. mrmre: an r package for parallelized mrmr ensemble feature selection. *Bioinformatics*, 29(18):2365–2368, 2013.

[DTP+07] Rebecca Dent, Maureen Trudeau, Kathleen I Pritchard, Wedad M Hanna, Harriet K Kahn, Carol A Sawka, Lavina A Lickley, Ellen Rawlinson, Ping Sun, and Steven A Narod. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical cancer research*, 13(15):4429–4434, 2007.

[Edw00] David Edwards. *Introduction to graphical modelling*. Springer Science & Business Media, 2000.

[EHC14] Soo-Heang Eo, Seung-Mo Hong, and HyungJun Cho. Identification of outlying observations with quantile regression for censored data. *arXiv preprint arXiv:1404.7710*, 2014.

[EKL06] T. Eltoft, T. Kim, and T. Lee. Multivariate scale mixture of Gaussians modeling. In *Independent Component Analysis and Blind Signal Separation*, pages 799–806. Springer Berlin Heidelberg, 2006.

[Far82] Vern T Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046, 1982.

[FB16] F. Fazayeli and A. Banerjee. *The Matrix Generalized Inverse Gaussian distribution: properties and applications*, volume 9851 of *Frasconi P., Landwehr N., Manco G., Vreeken J. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2016. Lecture Notes in Computer Science*. Springer, Cham., 2016.

[FCS+21] Jacques Ferlay, Murielle Colombet, Isabelle Soerjomataram, Donald M Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. Cancer statistics for the year 2020: An overview. *International journal of cancer*, 149(4):778–789, 2021.

[FGFBGM22] Laura Freijeiro-González, Manuel Febrero-Bande, and Wenceslao González-Manteiga. A critical review of lasso and its derivatives for variable selection under dependence among covariates. *International Statistical Review*, 90(1):118–145, 2022.

[FHT08] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics.*, 9(3):432–441, 2008.

[FKS20] Y. Fang, D. Karlis, and S. Subedi. A Bayesian approach for clustering skewed data using mixtures of multivariate normal-inverse Gaussian distributions. *arXiv:2005.02585*, 2020.

[Gir14] C. Giraud. *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2014.

[Gol84] Anne I Goldman. Survivorship analysis when cure is a possibility: a monte carlo study. *Statistics in Medicine*, 3(2):153–163, 1984.

[GSSS99] Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

[GYA18] Jonathan L Gross, Jay Yellen, and Mark Anderson. *Graph theory and its applications*. Chapman and Hall/CRC, 2018.

[GYNL19] L. Gan, X. Yang, N. Narisetty, and F. Liang. Bayesian joint estimation of multiple graphical models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[HCP+82] Frank E Harrell, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

[HJ12] R. A. Horn and C. R. Johnson. *Matrix Analysis (Second Edition)*. Cambridge University Press, Cambridge, New-York, 2012.

[HPH+21] Moritz Herrmann, Philipp Probst, Roman Hornung, Vindi Jurinovic, and Anne-Laure Boulesteix. Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in bioinformatics*, 22(3):bbaa167, 2021.

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, 2015.

[HW19] Roman Hornung and Marvin N Wright. Block forests: random forests for blocks of clinical and omics covariate data. *BMC bioinformatics*, 20(1):1–17, 2019.

[IKBL08] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, 2(3):841–860, 2008.

[Int] World Cancer Research Fund International. `https://www.wcrf.org/dietandcancer/breast-cancer-statistics/`. Accessed: 16-03-2022.

[JAB11] Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824, 2011.

[JMS+19] Yi-Zhou Jiang, Ding Ma, Chen Suo, Jinxiu Shi, Mengzhu Xue, Xin Hu, Yi Xiao, Ke-Da Yu, Yi-Rong Liu, Ying Yu, et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer cell*, 35(3):428–440, 2019.

[JOV09] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440, 2009.

[KC92] Anthony YC Kuk and Chen-Hsin Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541, 1992.

[KJH+18] Simon Klau, Vindi Jurinovic, Roman Hornung, Tobias Herold, and Anne-Laure Boulesteix. Priority-lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC bioinformatics*, 19(1):1–14, 2018.

[KK10] David G. Kleinbaum and Mitchel Klein. *Survival analysis*. Springer, 2010.

[KKK06] Yuwon Kim, Jinseog Kim, and Yongdai Kim. Blockwise sparse regression. *Statistica Sinica*, pages 375–390, 2006.

[KM58] Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[KWB18] Friedrich Kruber, Jonas Wurst, and Michael Botsch. An unsupervised random forest clustering technique for automatic traffic scenario categorization. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2811–2818. IEEE, 2018.

[KX12] Seyoung Kim and Eric P Xing. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *The Annals of Applied Statistics*, 6(3):1095–1117, 2012.

[Lau96] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[LBC+16] B. Liquet, L. Bottolo, G. Campanella, S. Richardson, and M. Chadeau-Hyam. R2GUESS: A graphics processing unit-based R package for Bayesian variable selection regression of multivariate responses. *J. Stat. Softw.*, 69(2):1–32, 2016.

[LL10] Caiyan Li and Hongzhe Li. Variable selection and regression analysis for graph-structured covariates with an application to genomics. *The annals of applied statistics*, 4(3):1498, 2010.

[LMC19] Z. Li, T. Mccormick, and S. Clark. Bayesian joint spike-and-slab graphical Lasso. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3877–3885. PMLR, 2019.

[LMPS17] B. Liquet, K. Mengersen, A. N. Pettitt, and M. Sutton. Bayesian variable selection regression of multivariate responses for group data. *Bayesian Anal.*, 12(4):1039–1067, 2017.

[LNZ15] Y. Li, B. Nan, and J. Zhu. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics.*, 71:354–363, 2015.

[Lok99] Justin Lokhorst. The lasso and generalised linear models. *Honors Project, The University of Adelaide, Australia*, 1999.

[Lu08] Wenbin Lu. Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics*, 60(3):545–574, 2008.

[Lu09] Z. Lu. Smooth optimization approach for sparse covariance selection. *Siam. J. Optimiz.*, 19(4):1807–1827, 2009.

[LW14] Elisa T. Lee and John Wang. *Statistical methods for survival data analysis*. John Wiley & Sons, 2014.

[MB06] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 34(3):1436–1462, 2006.

[MDLW18] M. Maathuis, M. Drton, S. L. Lauritzen, and M. Wainwright. *Handbook of Graphical Models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, 2018.

[MSH07] Shuangge Ma, Xiao Song, and Jian Huang. Supervised group lasso with applications to microarray data analysis. *BMC bioinformatics*, 8(1):1–17, 2007.

[MVDGB08] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[MW47] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[MW06] H. Massam and J. Wesołowski. The Matsumoto-Yor property and the structure of the Wishart distribution. *J. Multivariate. Anal.*, 97:103–123, 2006.

[OOJP21] Eunice Okome Obiang, Pascal Jézéquel, and Frédéric Proïa. A partial graphical model with a structural prior on the direct links between predictors and responses. *ESAIM: Probability and Statistics*, 25:298–324, 2021.

[OOJP22] Eunice Okome Obiang, Pascal Jézéquel, and Frédéric Proïa. A bayesian approach for partial gaussian graphical models with sparsity. *Bayesian Analysis*, 1(1):1–26, 2022.

[O'Q08] John O'Quigley. *Proportional hazards regression*, volume 542. Springer, 2008.

[PBL+10] E. Petretto, L Bottolo, S. R. Langley, M. Heinig, C. McDermott-Roe, R. Sarwar, M. Pravenec, N. Hübner, T. J. Aitman, S. A. Cook, and S. Richardson. New insights into the genetic control of gene expression using a bayesian multi-tissue approach. *PLOS Comput. Biol.*, 6(4):1–13, 2010.

[PBTB13] F. Pascal, L. Bombrun, J. Y. Tourneret, and Y. Berthoumieu. Parameter estimation for multivariate generalized Gaussian distributions. *IEEE. T. Signal. Process.*, 61(23):5960–5971, 2013.

[PC08] T. Park and G. Casella. The Bayesian Lasso. *J. Am. Stat. Assoc.*, 103(482):681–686, 2008.

[PCV15a] João Diogo Pinto, Alexandra M Carvalho, and Susana Vinga. Outlier detection in cox proportional hazards models based on the concordance c-index. In *International Workshop on Machine Learning, Optimization and Big Data*, pages 252–256. Springer, 2015.

[PCV15b] Joao Diogo Pinto, Alexandra M Carvalho, and Susana Vinga. Outlier detection in survival analysis based on the concordance c-index. In *BIOINFORMATICS*, pages 75–82, 2015.

[PD00] Yingwei Peng and Keith BG Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243, 2000.

[QXH+16] Jingdan Qiu, Xinying Xue, Chao Hu, Hu Xu, Deqiang Kou, Rong Li, and Ming Li. Comparison of clinicopathological features and prognosis in triple-negative and non-triple negative breast cancer. *Journal of cancer*, 7(2):167, 2016.

[Ra18] Agatha S Rodrigues and al. Use of interval-censored survival data as an alternative to kaplan-meier survival curves: studies of oral lesion occurrence in liver transplants and cancer recurrence. *Applied Cancer Research*, 38(1):1–10, 2018.

[RAM12] P. Rossi, G. Allenby, and R. McCulloch. *Bayesian Statistics and Marketing*. Wiley Series in Probability and Statistics. Wiley, 2012.

[RS06] J. Ramsay and B. Silverman. *Functional Data Analysis, 2nd ed.* Springer, New-York, 2006.

[RSZZ15] Z. Ren, T. Sun, C. H. Zhang, and H. H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Stat.*, 43(3):991–1026, 2015.

[RWRY11] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980, 2011.

[SFHT13] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.

[SK03] Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[SK12] K. A. Sohn and S. Kim. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics.*, volume 22 of *Proceedings of Machine Learning Research*, pages 1081–1089. PMLR, 2012.

[Sla12] M. Slawski. The structured elastic net for quantile regression and support vector classification. *Stat. Comput.*, 22:153–168, 2012.

[Soc] American Cancer Society. `https://www.cancer.org/cancer/breast-cancer/about/types-of-breast-cancer/triple-negative.html`. Accessed: 16-03-2022.

[Spe04] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[SSB+05] Tao Shi, David Seligson, Arie S Belldegrun, Aarno Palotie, and Steve Horvath. Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Modern Pathology*, 18(4):547–557, 2005.

[ST00] Judy P Sy and Jeremy MG Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

[SZCT10] M. Slawski, W. Zu Castell, and G. Tutz. Feature selection guided by structural information. *Ann. Appl. Stat.*, 4(2):1056–1080, 2010.

[TB06] Gerhard Tutz and Harald Binder. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62(4):961–971, 2006.

[TCJL+10] Aye Aye Thike, Poh Yian Cheok, Ana Richelia Jara-Lazaro, Benita Tan, Patrick Tan, and Puay Hoon Tan. Triple-negative breast cancer: clinicopathological characteristics and relationship with basal-like breast cancer. *Modern pathology*, 23(1):123–133, 2010.

[TGF90] Terry M Therneau, Patricia M Grambsch, and Thomas R Fleming. Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160, 1990.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[Tib97] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[TSZY17] Zaixiang Tang, Yueping Shen, Xinyan Zhang, and Nengjun Yi. The spike-and-slab lasso cox model for survival prediction and associated genes detection. *Bioinformatics*, 33(18):2799–2807, 2017.

[UCP+11] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[VDWLV+16] Mark A Van De Wiel, Tonje G Lien, Wina Verlaat, Wessel N van Wieringen, and Saskia M Wilting. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in medicine*, 35(3):368–381, 2016.

[W+01] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[Wai09] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[Wer76] Nanny Wermuth. Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics*, pages 95–108, 1976.

[Whi09] Joe Whittaker. *Graphical models in applied multivariate statistics*. Wiley Publishing, 2009.

[Wil45] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[WJ63] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[WRHG20] R. Wei, B. J Reich, J. A. Hoppin, and S. Ghosal. Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures. *Statist. Sinica*, 30:55–79, 2020.

[XG15] X. Xu and M. Ghosh. Bayesian variable selection and estimation for Group Lasso. *Bayesian Anal.*, 10(4):909–936, 2015.

[XSM+16] Z. Xu, D. F. Schmidt, E. Makalic, G. Qian, and J. L. Hopper. Bayesian grouped horseshoe regression with application to additive models. In *AI 2016: Advances in Artificial Intelligence*, pages 229–240. Springer International Publishing, 2016.

[YL06] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[YN20] X. Yang and N. Narisetty. Consistent group selection with bayesian high dimensional modeling. *Bayesian Anal.*, 15(3):909–935, 2020.

[Yur83] Nesterov Yurii. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

[YZ14] X. T. Yuan and T. Zhang. Partial Gaussian graphical model estimation. *IEEE. T. Inform. Theory.*, 60(3):1673–1687, 2014.

[ZCa16] Jianfei Zhang, Lifei Chen, and al. Survival prediction by an integrated learning criterion on intermittently varying healthcare data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[ZRY09] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

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

Titre : Contribution à l'étude des modèles graphiques et statistique en grande dimension appliquée à la modélisation du cancer du sein triple négatif.

Mot clés : régression linéaire en grande dimension, modèle graphique partiel, corrélation partielle, pénalisation structurante, sparsité, approche bayésienne, spike-and-slab, échantillonneur de Gibbs, sélection de variables.

Résumé : Cette thèse s'articule autour de deux axes. Le premier constitue une contribution à l'étude des modèles graphiques gaussiens partiels (PGGM) dans le cadre de l'apprentissage en grande dimension. Plus précisément, nous nous intéressons à la modélisation à sorties multiples, où nous souhaitons estimer d'une part la matrice Δ des liens directs entre les prédicteurs et les réponses, et d'autre part la matrice de précision conditionnelle des réponses Ω_y. Nous débutons avec une approche fréquentiste par maximum de vraisemblance pénalisée, où nous proposons un PGGM muni de deux formes de pénalisation : une pénalisation ℓ_1 induisant de la sparsité sur Δ et Ω_y, et une pénalisation structurante reflétant un a priori gaussien généralisé sur les liens directs. Nous montrons que, lorsqu'il est convenablement régularisé, ce modèle est agrémenté d'une garantie théorique prenant la forme d'une borne supérieure sur l'erreur d'estimation. Enfin, nous clôturons cette première réflexion par des études empiriques mettant en avant le caractère structurant de cette procédure d'estimation, et sa pertinence sur un jeu de données réelles. Nous poursuivons par l'étude de la contrepartie bayésienne, jusqu'alors inexplorée dans la littérature. En suivant une stratégie spike and slab, nous offrons plusieurs structures hiérarchiques imposant soit une configuration saturée, sparse, group-sparse ou encore sparse-group-sparse de la matrice Δ. Nous obtenons une garantie théorique pour les configurations sparse et group-sparse, et illustrons les résultats compétitifs de ces modèles sur une étude de simulation et un jeu de données réels, menés avec des échantillonneurs de Gibbs. Le deuxième axe de la thèse est, quant à lui, entièrement dévolu à la sélection de variables pronostiques en analyse de suvie multi-omique. Nous y proposons un algorithme de sélection de variables descendante offrant un consensus entre différentes méthodes de régularisation, notamment celles présentées dans le premier axe. L'efficacité de cette approche est enfin étudiée sur des données relatives au cancer du sein triple négatif, en prenant le soin de répondre aux contraintes identifiées par les oncologues. Tous nos codes sont rendus disponibles à la communauté.

Title: Contribution to the study of graphical models and high-dimensional statistics applied to the modeling of Triple-Negative Breast Cancer.

Keywords: high-dimensional linear regression, partial graphical model, partial correlation, structural penalization, sparsity, Bayesian approach, spike-and-slab, Gibbs sampler, variable selection.

Abstract: This thesis is articulated around two axes. The first one is a contribution to the study of partial Gaussian graphical models (PGGM) in high-dimensional learning. Precisely, we are interested in the multiple-output modeling, where we aim at estimating, on the one hand the matrix Δ of direct links between predictors and responses, and on the other hand the conditional precision matrix Ω_y of responses. We start with a frequentist approach by penalized maximum likelihood, where we propose a PGGM with two forms of penalization: a ℓ_1 penalty inducing sparsity on Δ and Ω_y, and a structural penalty reflecting a generalized Gaussian prior on the direct links. We show that, when properly regularized, this model comes with a theoretical guarantee taking the form of an upper bound on the estimation error. Finally, we close this first reflection with empirical studies highlighting the structuring property of this estimation procedure, and its relevance on a real dataset. We continue with the study of the Bayesian counterpart, previously unexplored in the literature. Following a spike and slab strategy, we offer several hierarchical structures imposing either a saturated, sparse, group-sparse or sparse-group-sparse configuration of the matrix Δ. We obtain a theoretical guarantee for the sparse and group-sparse configurations, and illustrate the competitive results of these models on a simulation study and a real dataset, conducted with Gibbs samplers. The second part of the thesis is entirely devoted to the selection of prognostic variables in multi-omics survival analysis. We propose a stepwise variable selection algorithm offering a consensus between different regularization methods, including those presented in the first axis. The efficiency of this approach is finally studied on a dataset relating to triple-negative breast cancer, while taking care to meet the constraints identified by oncologists. All our codes are made available to the community.