Contributions to variable selection in high-dimension and its uses in biology

Perrine Lacroix

Contributions to variable selection in high-dimension and its uses in biology

Contributions à la sélection de variables en grande dimension et ses utilisations en biologie

Thèse de doctorat de l'université Paris-Saclay

École doctorale de mathématiques Hadamard n°574 (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques, Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche Laboratoire de mathématiques d'Orsay (Université Paris-Saclay, CNRS) et Institute of Plant Sciences Paris-Saclay (IPS2) (Université Paris-Saclay, CNRS, INRAE, Univ Evry), sous la co-direction de Pascal MASSART, professeur, et de Marie-Laure MARTIN, directrice de recherche.

Thèse soutenue à Paris-Saclay, le 16 Décembre 2022, par

Perrine LACROIX

Composition du jury

Membres du jury avec voix délibérative

Gilles BLANCHARD Professeur, Laboratoire de Mathématiques d'Orsay - Université Paris-Saclay	Président
Adeline LECLERCQ SAMSON Professeure, Laboratoire Jean Kuntzmann - Université Grenoble Alpes	Rapporteure & Examinatrice
Pierre NEUVIAL Directeur de recherche, Institut de Mathématiques de Toulouse - Université de Toulouse	Rapporteur & Examinateur
Claire LACOUR Professeure, Laboratoire d'analyse et de mathématiques appliquées - Université Gustave Eiffel	Examinatrice
Franck PICARD Directeur de recherche, Laboratoire de Biologie et Modélisation de la Cellule - ENS Lyon	Examinateur

THESE DE DOCTORAT

NNT : 2022UPASM039

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre : Contributions à la sélection de variables en grande dimension et ses utilisations en biologie

Mots clés : grande dimension, sélection de variables, calibration de pénalités, risque prédictif et taux de fausses découvertes, heuristique de pente, identification de gènes

Résumé : La révolution des données que nous connaissons aujourd'hui se caractérise par la prolifération de données massives dans tous les domaines d'activités économiques, mais aussi dans les sciences. Cette révolution des données scientifiques concerne en particulier la biologique moléculaire. L'étude de l'expression des gènes d'un organisme est l'exemple clé mis en avant dans cette thèse. Les données d'expression de gènes sont typiquement caractérisées par un nombre élevé de variables descriptives pour un nombre d'observations restant limité. Identifier les variables pertinentes constitue une étape cruciale pour l'exploitation des données ainsi que leur interprétation. Cette thèse est centrée sur la question de la sélection de variables dans le cadre statistique de la régression linéaire gaussienne en grande dimension. Le cœur de notre analyse repose sur l'introduction de nouvelles fonctions de pénalité pour le critère d'ajustement des moindres carrés. Celles-ci dépendent de constantes, que nous voyons comme des hyperparamètres à calibrer sur le jeu de données d'étude. L'originalité de notre approche réside en l'introduction du *False Discovery Rate (FDR)* pour réaliser cette calibration. Dans un premier temps, nous prouvons un encadrement théorique du FDR lorsque les variables sont ordonnées, puis nous mettons en place un algorithme de calibration de l'hyperparamètre pour satisfaire un compromis entre le contrôle du risque prédictif et celui du FDR. Pour sélectionner des variables non ordonnées en grande dimension, nous revisitons le thème de la sélection de variables via la minimisation d'un critère convexe de type *Lasso*. Nous proposons une approche qui consiste à choisir les variables, ordonnées par le *chemin de régularisation*, via une méthode de pénalisation adaptative. Des simulations intensives mettent en évidence l'intérêt du rééchantillonage et des pénalités non-asymptotiques. Nous généralisons la méthode de calibration adaptative de pénalité dite "de l'*heuristique de pente*" à la calibration de deux hyperparamètres simultanément ainsi qu'au contexte d'une collection de modèles aléatoires qui est ici le nôtre. Enfin, notre nouvel algorithme, ainsi que certaines procédures de sélection de variables, sont appliqués sur un jeu de données transcriptomiques d'*Arabidopsis thaliana*. L'identification des *facteurs de transcription* de gènes cibles constitue ici la problématique biologique.

Title : Contributions to variable selection in high-dimension and its uses in biology

Keywords : high-dimension, variable selection, penalty calibrations, predictive risk and false discovery rate, slope heuristics, gene identification

Abstract : The current data revolution is characterized by the proliferation of massive data in all areas of economic activity and in sciences as well. In particular, this scientific data revolution concerns molecular biology. The study of gene expression in an organism is the key example in this thesis. Gene expression data are typically characterized by a high number of descriptive variables for a limited number of observations. Identifying the relevant variables is a crucial step for data exploitation and interpretation. This thesis focuses on the issue of variable selection in the statistical framework of high-dimensional Gaussian linear regression. The core of our analysis is based on introducing new penalty functions for the least squares adjustment criterion. These penalties depend on constants, considered as hyperparameters to be calibrated on the available data set. The originality of our approach lies in introducing the *False Discovery Rate (FDR)* to perform this calibration. First, we establish theoretical lower and upper bounds of the FDR when the variables are ordered. Then we set up a calibration algorithm of the hyperparameter to satisfy a trade-off between the predictive risk control and the FDR one. To select non-ordered variables in the high-dimension setting, we revisit the topic of variable selection via the minimization of a convex criterion such as the *Lasso*. We propose an approach to select the variables, which are ordered by the *regularization path*, via an adaptive penalization method. Intensive simulations show the interest in resampling and in non-asymptotic penalties. We generalize the adaptive penalty calibration method called "*the slope heuristics*" to calibrate two hyperparameters simultaneously and in the context of a random model collection. Finally, our new algorithm and some variable selection procedures are applied to a transcriptomic dataset of *Arabidopsis thaliana*. The biological problem here consists in identifying *transcription factors* of target genes.

À mon Papi, qui aurait sans doute aimé lire cette thèse.

À ma Mamie, mon modèle de force et de courage.

À la famille qui s'agrandit.

Remerciements

Parce que sans vous, cette thèse n'aurait pas eu lieu, mes premiers remerciements s'adressent à mes directeur/trice de thèse, Pascal et Marie-Laure. Marie-Laure, merci de m'avoir fait confiance en me proposant ce sujet. Il m'a animée pendant plus de trois ans et me passionne toujours autant. Tes connaissances en biologie et ton positionnement inter-disciplinaire m'impressionnent et je te remercie sincèrement de m'avoir appris à me situer à l'interface entre les statistiques et la biologie. Merci pour toutes les discussions mathématiques que je n'oublierai pas. Pascal, ma plus grande victoire de ces trois années aura été le jour où tu m'as renommée « le microbe » : yes ! J'aurais réussi à taquiner un grand bonhomme comme toi. Je ne sais pas si je t'ai mis des « paillettes dans les yeux », en tout cas, toi tu en as mis dans les miens. Je te remercie pour les discussions scientifiques mais aussi les non scientifiques (qui m'ont bien aidée !) et ce, toujours accompagnées d'un café (allongé bien sûr !). Promis, à chaque futur sifflement dans n'importe quel couloir de n'importe quel lieu, j'aurais une pensée pour toi. A tous les deux, merci pour cette chance, pour votre patience, vos encouragements et vos conseils.

Je remercie chaleureusement Adeline Leclercq-Samson et Pierre Neuvial de m'avoir fait l'honneur de rapporter cette thèse. Je suis extrêmement reconnaissante de vos relectures attentives et approfondies. Merci pour les pistes de réflexion, les encouragements et les compliments que vous m'avez communiqués dans vos rapports. J'adresse également mes sincères remerciements à Claire Lacour et Franck Picard d'avoir eu la gentillesse d'être membres de mon jury de thèse et Gilles Blanchard d'avoir accepté de le présider.

Je remercie affectueusement mes deux équipes, GNet et Proba-Stat, pour l'accueil et la complicité.
Je dois à l'IPS2 tout ce que j'ai appris en biologie, en informatique (enfin... j'aurais toujours besoin de toi Jean-Philippe, tu n'es pas débarrassé !) et en bio-statistique. Merci Etienne et Benoît, mes biologistes préférés (j'omettrai de mettre un ordre pour ne vexer personne). Merci Etienne de te plonger dans les statistiques et d'y apporter ton expertise biologique. Benoît, ma limite inférieure au foot (booooouuuhh !), sache que ton animal mort m'a accompagnée de nombreux mois, tel un porte-bonheur. Tes vacheries m'ont rendue plus forte, je t'en suis reconnaissante. Je suis fière d'avoir été une Ginette. J'y ai reçu de belles leçons de vie. Cécile, je te remercie de m'avoir transmis ton sens de l'organisation, de planification, du rangement, du tri. Grâce à toi, le « je garde, au cas où » s'est transformé en « je ne m'en sers plus ça. Où ça se recycle ? Il faut lui donner une seconde vie, viiiiiite ! ». Je suis encore loin de ton niveau mais je vais essayer de toujours m'en approcher. Véro, ta joie de vivre nous donne envie de nous lever le matin. Merci pour tes rires communicatifs ! Merci Guillem et Arnaud pour l'intérêt porté à mon sujet de thèse et pour vos questions pertinentes. Christine, merci de t'être toujours préoccupée de savoir si j'allais bien, surtout pendant la période de rédaction. Ton attention m'a beaucoup touchée. Margot, mon binôme Verseau (bon OK.. on est un trio chez les Ginettes. OUPS !) merci pour les taquineries partagées. Merci Nathalie, Marie-Hélène,

Yacine, Simon et les anciens : Lamia et Nadia. Kévin, tu apprendras par ces lignes qu'on est toujours capable de s'allier à ses pires ennemis pour arriver à ses fins: ta date d'anniversaire, c'est Jean-Philippe qui me l'a balancée ! Merci à toi, pour tout ! Pour ta bonne humeur, ta bienveillance et ta sensibilité, pour tes piques et tes attaques. Merci de m'avoir laissée me confier à toi, alors même que tu vis à des kilomètres. Merci Julien d'avoir été mon jumeau dans cette aventure de thèse. Ensemble, on aura stressé, encaissé les échecs, répondu aux échéances mais aussi savouré les réussites, les périodes plus zen et le bonheur de se voir avancer. A deux, c'est quand même bien mieux que tout seul ! Jean-Philippe (ou JPhT, Philippe ou Phiphi [1], que préfères-tu ??), je crois que je te dois mille excuses. Pardon pour les longs mois où tu as du me supporter. Je suis ravie de t'avoir fait sortir de tes gonds plusieurs fois, c'est un vrai régal d'avoir plus de répartie que toi ! On aura bien rigolé tous les deux ! Grâce à toi, j'ai appris à bien doser les cafés mais aussi qu'il était possible d'accepter autant de compliments pour des compotes (les meilleures que je n'ai jamais mangées, n'hésite pas à m'en redonner !) alors que tu ne fais que la découpe et la coupe des pommes. Je terminerai par : tu as été mon premier punching-ball, mais... tu me connais... tu sais que je ne pense pas une seule de mes excuses et que s'il fallait recommencer, je t'attaquerais exactement comme je l'ai fait. Plus sérieusement, merci d'avoir toujours répondu présent, d'avoir endossé le rôle de psy plus d'une fois. J'ai pu me reposer sur ton soutien et ça, je ne l'oublierai pas.
Je remercie l'équipe Proba-Stat pour les moments de convivialité au CESFO ou lors des nombreux cafés en 3P15. Là encore, les murs auront entendu mon rire à de multiples reprises. Nico, tu as été la première personne à qui j'ai parlé, d'abord dans ton bureau au 425 où tu m'as arraché quelques larmes, puis en 3P15, salle dont j'ignorais l'existence jusqu'alors. Je devrais te remercier pour l'intégration dans l'équipe mais je sais que tu n'as fait qu'agir sous les ordres de Camille. Alors, immense merci Camille de porter autant attention aux gens qui t'entourent. Je serais passée à côté de plein de belles choses si tu n'avais pas été là. Je sais que je peux compter sur ton soutien en tant que chercheuse en maths pour la bio et en tant que femme. Guillermo, j'ai eu la chance de bénéficier de ton expertise sur les tests multiples. Merci à toi ! J'ai appris plein de choses grâce à tes multiples (comme les tests ! Oups, blague nulle, on dirait Nico...) corrections. Merci de m'avoir donné de ton temps. Je suis ravie de faire partie de tes 800 amis, je crois que c'est mérité après tant de petits-déjeuners forcés partagés avec toi à Fréjus. J'espère qu'on aura l'occasion de se revoir lors de ton prochain anniversaire mais j'ai peur d'être *emportée par la foule*. Nico, en tant que semi-paramétrique, j'espère avoir bien pourri ton existence, mais rassure toi, Olivier en a pris aussi pour son grade (encore Joyeux Anniversaire Olivier !). Merci à \cite{Naulet2022} pour m'avoir ouvert les yeux sur ma naïveté : évidemment, ce manuscrit n'a pas une pointe de valeur sans une citation de toi. C'est chose faite. Comme ce sont des remerciements, merci, Zach, pour " $\cdot$ " . Merci au groupe *quatre jeunes et un vieux* : Guillaume, Alice, Etienne et Jérome. La vie aurait été moins belle sans vous. Je garde un souvenir mémorable des vacances au ski (hop là !), du week-end à la Tranche (les broco-truites !), des matchs de basket ou encore des soirées au HoB. Merci de m'avoir supportée à Côme, je sais que je n'ai pas toujours été facile. Jérome, il est grand temps que je parte, tu commences à me lancer beaucoup trop de piques (d'ailleurs, si j'avais été toi,

[1] surnoms donnés respectivement par tes collègues, ta famille, tes amis de plongée. Je balance là non ?

je n'aurais jamais joué pique...). Vous avez été formidables !! Merci à Etienne, Alice, Vincent, Jérome et Nico pour vos passages au LMO lors de mes week-ends de rédaction pour un repas, un café, une discussion, un soutien ou un simple coucou. Louis, j'adore t'embêter mais tu me le rends tellement bien aussi. On aura emprunté plusieurs routes ensemble, à vélo pendant le Covid, ou sur le chemin de la thèse (et oui, tu remarques l'effort de « méthaphor(a)e »). Merci pour les moments de complicité. Gilles, ta gentillesse est à ta hauteur ! Tu as toujours pris soin de t'assurer que j'allais bien, merci ! Merci à *Adam team* (« où il est le bébé ?? ») pour les tortures calculatoires, merci au groupe Pizzama pour les brillant exposés qui y ont été donnés. Merci à Jean-Baptiste. L, Elisabeth, Zacharie, Rémi et Arnaud pour l'encadrement lors du monitorat. Vos conseils pédagogiques sont riches en enseignement (!). Elisabeth, je lève mon chapeau à ta grande disponibilité et ton investissement à l'enseignement malgré ton emploi du temps chargé. C'était un très grand plaisir de travailler à tes côtés et j'essaierai à l'avenir de reproduire l'exemple que tu incarnes. Zacharie, c'est ton efficacité que je garderai en tête. Malgré ce que tu laisses croire, tu n'as rien laissé de côté et c'est très agréable de faire équipe avec toi. Sylvain, je t'ai souvent embêté avec mes questions. Je te remercie affectueusement pour toute la biblio vers laquelle tu m'as orientée, pour toutes les pistes de réflexion que tu m'as offertes et pour les discussions enrichissantes à la fin de mes exposés. Merci de m'avoir fait confiance en me proposant de co-organiser le séminaire des élèves. J'ai appris plein de choses, découvert plein de domaines, rencontré plein de gens, bref, j'ai adoré l'expérience ! Merci Vincent. D, Wojciech, Matthieu et Gilles de m'avoir accompagnée. En particulier, merci Matthieu de nous avoir toujours laissé galérer avec les problèmes techniques, on en ressort plus fort ! Et merci Wojciech, j'ai toujours pu compter sur toi et tu as toujours su calmer mes moments de stress. Je n'oublie pas Nathanael, Paul, Jean-François, Liliane, Jean-Michel, Edouard, Arvind, Rémi, Armand, Olympio, Bastien, Jean-Baptiste.F, Jérémie, Simon, Samy. A tous, je vous dois les plus sincères des remerciements. Les derniers mois de thèse ont été les plus riches en émotions mais votre soutien a été sans faille. Je vous souhaite une grosse fiesta à mon départ (chez Guillermo?). Soyez simplement prudents sur le ventriglisse, certains y ont laissé des côtes.

Je souhaite profiter de ces remerciements pour lever mon chapeau à l'ensemble des enseignants que j'ai pu croiser sur mon chemin. C'est par votre pédagogie et votre partage de connaissances que j'ai pu en arriver là aujourd'hui. Le premier à m'avoir tirée vers le haut est mon professeur de maths de cinquième, merci. J'ai une tendre pensée pour Mme C. Lepez. J'envie l'énergie, la passion et l'enthousiasme que vous transmettez chaque jour à vos élèves. Vous faites partie de mes modèles de femme. Je mesure la chance d'être arrivée à Orsay pour le magistère. Je remercie les enseignants, en particulier David Harari, Dominique Hulin, Elisabeth Gassiat, Édouard Maurel-Segala, Pierre-Guy Plamondon, Laurent Moonens, Sophie Lemaire et tous ceux que je n'ai pas cités. Merci Frédéric Paulin pour votre implication auprès des étudiants du magistère. Merci Nathalie Carrière pour ta bonne humeur et ta présence. Je remercie Sylvie Méléard, Christophe Giraud, Liliane Bel, Stéphane Robin, Estelle Kuhn, Paul-Henry Cournède, Jean-Philippe Vert, pour votre enseignement d'excellence et les discussions scientifiques au sein du M2 MSV. Christophe, tu es le premier a m'avoir appris ce que sont les statistiques en grande dimension. Ton livre m'a accompagnée de nombreuses fois pendant cette thèse, il m'est d'une grande aide, merci à toi. Liliane, merci de m'avoir recommandée dès la fin de mes études et de ne pas m'avoir lâchée et ce, jusqu'au jour de la soutenance. Merci

à mes encadrants de projets et de stage : Claude Zuily, Loïc Rodin, Claire Lacour, Etienne Delannoy, Mélina Gallopin et Marie-Laure. Vous avez été témoins de mes premiers pas dans la recherche. Je remercie affecteusement Stéphane Nonnenmacher et Hans Rugh que j'ai sollicités à de nombreuses reprises pour des questions administratives sur la thèse, sur l'agrégation, sur les prolongement possibles, etc... Merci pour votre patience, votre présence et pour le travail effectué afin que tout se passe toujours au mieux. Ce paragraphe ne pouvait se terminer qu'en adressant ma reconnaissance immuable envers Patrick Gérard. Je vous remercie pour l'intérêt permanent que vous avez eu à mon égard du premier jour de M1 jusqu'au jour de ma soutenance en passant par l'agrégation. Je n'oublierai jamais votre soutien qui m'a fait tenir pendant la période de deuil. Merci du fond du coeur.

C'est en partie grâce aux discussions que j'ai pu mûrir sur mon sujet de thèse. Je remercie en particulier Mélina Gallopin pour ton encadrement pendant le stage pré-thèse. Ton expertise m'a permis d'avancer; tes conseils et astuces de code sur R m'ont rendue autonome et à l'aise. Merci Cathy Maugis pour l'intérêt que tu as manifesté vis à vis de mon sujet de thèse. Tu as toujours répondu présente pour discuter, par mails ou de visu et j'en suis toujours sortie avec une longue fiche de notes et d'idées. Samir Dou, c'était un plaisir de découvrir ton travail et d'échanger des discussions scientifiques avec toi. Enfin, merci à Marion Naveau et Armand Favrot, les premiers étudiants que j'ai encadrés. Vous m'avez fait confiance et accompagnée dans l'étude des données réelles, et ce malgré les conditions sanitaires. Vous m'avez donné envie de renouveler l'expérience d'encadrement.

« Il y a les amis... »
Pour votre soutien inconditionnel et parce que ça fait des années que nos vies se croisent, je remercie les *Champions d'Agreg et du monde* : Julie, Julien, Christian, Fabien, Valentin, Arnaud, Laurène, Florian et Manu. Fabien, à ton tour maintenant ! Vous voir grandir, vous épanouir et être témoins de vos débuts de vie d'adulte est un réel cadeau. Une mention spéciale à Julie car tu veilles en permanence au bien-être de chacun et à la cohésion du groupe. Merci de casser ta carapace quand tu es avec nous. Tu es une perle. Lucile & Antoine, merci pour les rires et la déconnade. *Je n'ai pas d'amis (,) comme vous*, et là, « y'a deux écoles » pour la virgule. Antoine, « ce que tu dis me met mal à l'aise ». Lucile, merci pour ton humour et ta gentillesse ! Merci mes chers voisins Guillaume et Amandine, pour les plantes, les jeux de société et les cadeaux canadiens. Merci Florian Lasgorceux pour la confiance que tu m'accordes depuis la prépa. Merci les copains du rock et de la danse africaine pour les bols d'air, de détente et de rire. Cassandra, Maxime, Alexandre, la distance nous éloigne mais l'amitié, c'est aussi *être séparés et que rien ne change.*

« ... Il y a la famille... »
Maman, tu es exceptionnelle, je t'aime inconditionnellement. C'est grâce à ta disponibilité, ton implication dans la scolarité, ton aide depuis toute petite que j'ai pu gravir les échelons. Je mesure la hauteur de ton sacrifice. Cette réussite, je te la dois en grande partie. Merci d'être une maman incroyable. Papa, merci d'avoir toujours répondu présent à mes besoins. Tes aller-retours Lille-Orsay m'ont aidée à passer le cap de l'indépendance. Je sais que je peux toujours compter sur ton soutien et ton aide. Merci Renaud, Elodie et Aubin, la team fraternelle. Grandir auprès de vous est une chance et à de nombreuses reprises, vous m'avez

prouvé qu'on pouvait toujours compter les uns sur les autres. Elodie, même combat en ce mois de Décembre : « une soutenance, c'est comme un accouchement mais il n'y a pas besoin d'allaiter le manuscrit et à la fin, on a le droit de picoler [2]». Tu offres à la famille le plus beau cadeau. J'endosse un nouveau rôle, celui d'être Tata. Je n'ai aucune idée de comment agir mais je sais que ce sera riche en émotion. Merci Damien de changer nos vies. Merci Manon, à nous deux, on fait deux belles pestes. J'en profite pour remercier la grande famille. Malgré la distance, c'est toujours un bol d'air pur que de rentrer se ressourcer auprès de vous. Emma, tu as toujours cru en moi, merci pour ton soutien permanent et le temps que tu m'as donné lors de nos nombreux échanges téléphoniques. Merci Laëtitia d'être la Marraine que tu es, pour ta présence, année après année. La belle famille, merci pour votre soutien indéfectible. Vous n'avez cessé de prendre des nouvelles de cette thèse, de me rassurer, de me consoler. Beau Papa, merci d'avoir encaissé toutes mes attaques. Même si j'ai du mal à le reconnaître, on a beaucoup de points communs toi et moi. Belle Maman, merci de toujours avoir les pieds sur terre. Tes conseils sont souvent les meilleurs et c'est très rassurant de savoir que je peux compter sur toi. Beau frère, comment j'aurais pu faire sans tes cahiers de note en cadeaux ? Malgré tout, j'aimerais quand même savoir : quand partirons-nous en vacances ensemble ?

« ... Et puis il y a les amis qui deviennent une famille. » *Le groupe de soutien.*
Nicolas, la pépite de l'équipe. Quelle énergie, quel peps, quelle bonne humeur au quotidien ! Tu es impressionnant ! Merci pour ton entrain, tes blagues pas drôles et tes idées loufoques. Je retiens de toi la picole, le gospel (wohoo !), le jury matheux qui délivre les diplômes de M2, la picole, les karaokés, la (non) biblio, les barbecues, la picole. Tu nous auras bien fait rire ! Merci d'être toi, tu es parfait !!
Vincent, aucune ligne ne pourra décrire la hauteur du respect et de la gratitude que j'éprouve pour toi, ni tout ce que je te dois. Tu comptes énormément pour moi et je dois dire que je suis en totale admiration devant ta gentillesse infinie et ta personnalité si touchante. Il n'y a pas de merci assez fort mais je vais quand même essayer de t'en dire quelques uns... Je te dois mille mercis pour les mille rôles que tu as joués (faut dire, tu es un excellent acteur !). Je remercie le collègue que tu es. C'est un vrai régal de travailler avec toi, sur mes sujets comme les tiens. Tu m'as toujours considérée d'égal à égal. Scientifiquement, tu es précieux, et je mesure la chance que j'aie de pouvoir continuer l'aventure avec toi. Je remercie le (petit) camarade que tu es. Tu as accepté de chercher le σ^2 avec moi malgré la source de stress que cette erreur a engendrée. Je remercie le Papa que tu es. Tu veilles en permanence à ce que je garde (un peu) confiance en moi et je sais que ça te demande des heures de travail. Je remercie l'ami que tu es. Merci à tes épaules, qui supportent tous mes maux et merci pour les moments de joie à travers les répet's théâtre, nos pipeletteries ou encore notre mousse au chocolat mémorable. Tu m'as fait confiance, surtout ne changeons rien de nos habitudes ! Merci la vie de t'avoir mis sur ma route.
Alice, ma grande saucisse. Tu es une patate qui en vaut 7 (et ça c'est parfait pour la raclette !). Je t'écris ces mots bougies allumées, tisane dans la main et le coeur serré. Je n'ai plus jamais connu l'ennui depuis notre rencontre grâce aux nombreuses premières fois que j'ai eues avec toi : la planche, les ballets, le surf, l'orchestre, les friperies, le cata. Je n'oublierai jamais nos

[2] V.Rivoirard, 23/08/22

soirées mamies sur fond de musique des années 80. Merci pour ton soutien sans faille, pour m'avoir consolée plusieurs fois, bousculée quand il le fallait et surtout merci pour la lourdeur de ton humour. Tu es la vraie définition de l'amitié, d'ailleurs « qui trouve un ami, trouve un trésor. Merci ! Merci pour rien, merci pour tout ». Etienne a raison : « ne change rien, tu es précieuse ».
Etienne, mon pilier, mon bras droit, mon bras gauche (heu...). On s'est soutenu n'importe quand, n'importe où, pour n'importe quelle raison et sans relâche. Tu m'as accompagnée dans chacune de mes épreuves et je t'en remercie car je sais que mes souffrances sont tes souffrances. Je suis certaine que ces trois années nous ont rendus plus forts. Ta présence, ton soutien, ta générosité, ta patience me sont une aide incommensurable, à la hauteur de ce que tu représentes pour moi. Merci de jouer le NeuNeu pour me faire rire quand il faut, tu es le soleil dans l'obscurité. Merci d'être toi, merci d'avoir crée un nous. C'est avec toi que je grandis et quand je regarde derrière moi, je sais que tu es en grande partie responsable de la femme que je suis aujourd'hui. T'avoir à mes côtés est une chance incroyable et j'ai hâte de découvrir ce qui sera mis sur notre chemin.

Contents

Chapter 1

Introduction

Depuis quelques années, la statistique est amenée à répondre aux défis posés par le développement des technologies de pointe, des moyens de stockage des données et des avancées informatiques qui encouragent la création de bases de données de plus en plus massives et riches en information. Cette révolution des données concerne en particulier la biologie, discipline pourvoyeuse de grands jeux de données (complexes).

Les nouvelles bases de données sont caractérisées par un nombre croissant de variables descriptives pour un nombre d'observations restant constant. Les données sont dites de grande dimension. C'est le cas dans de nombreux domaines : la médecine, l'astrophysique, les sciences sociales, la finance, l'administration ou encore la biologie. L'une des particularités de ces données est l'acquisition d'observations simultanées de centaines, de milliers voire de millions de variables. Par exemple, dans le cadre de la biologie moléculaire, les technologies à haut-débit mettent à disposition des milliers de données transcriptomiques qui permettent de mesurer simultanément l'expression de tous les gènes d'un organisme. Cela permet de comprendre des phénomènes à l'échelle globale comme par exemple les mécanismes de régulation des organismes vivants. Il est naturel de penser que l'abondance des variables dans un jeu de données est un atout pour comprendre un phénomène ou pour identifier des interactions entre des entités par exemple. Mais l'analyse de données de grande dimension est extrêmement difficile. De plus, généralement, seules quelques unes des variables disponibles suffisent à répondre à la problématique posée: elles contiennent toute l'information nécessaire. Les autres variables peuvent être inutiles car redondantes, non pertinentes voire même nuisibles à l'étape d'estimation statistique. Identifier les variables pertinentes est incontournable pour obtenir une estimation de qualité mais aussi pour faciliter l'interprétation des praticiens qui préfèrent comprendre un phénomène à travers un petit nombre de variables.

Toujours plus nombreuses, ces données présentent également de plus en plus d'hétérogénéité, prennent des structures complexes et sont de plus en plus inter-dépendantes (structure spatiale, hiérarchique, lien de cause à effet,...). C'est le cas des données sous forme de textes, d'images ou vidéos, d'arbres, ou encore de graphes.

En régression linéaire gaussienne en grande dimension, une profusion de méthodes de sélection

de variables est disponible. D'un point de vue applicatif, ce modèle rencontre un succès certain en génomique. Ainsi, par exemple pour trouver les marqueurs moléculaires d'une maladie, on régresse le statut du patient sur l'expression de ses gènes [Golub et al., 1999]. Un autre exemple est la construction des réseaux d'interactions géniques. L'expression d'un gène est régressé sur l'expression de tous les autres gènes. Cet engouement a permis le développement de plusieurs packages R (huge, glmnet,...) qui permettent de mettre en oeuvre facilement ce type d'approche sur un jeu de données réelles. D'un point de vue méthodologique, de nombreux défis restent encore à relever.

Cette thèse se focalise sur la sélection de modèles via les chemins de régularisation. L'objectif principal des travaux exposés dans ce manuscrit consiste à proposer des développements théoriques et algorithmiques de pénalités calibrées directement sur le jeu de données disponibles, la finalité étant l'application sur le jeu de données réelles pour retrouver les facteurs de transcription d'un gène cible chez la plante *Arabidopsis thaliana*.

Contents

1.1 Problématiques biologique et statistique

1.1.1 La régulation de gènes par les *facteurs de transcription*

En biologie moléculaire, l'un des centres d'intérêts majeurs est l'amélioration de la connaissance sur le fonctionnement du génome des organismes. Dans cette thèse, nous nous intéressons à la plante *Arabidopsis thaliana* (Figure1.1). Cette plante fait partie de la famille des *brassicacées* qui regroupe essentiellement des plantes herbacées de l'hémisphère Nord. Depuis le début des années 2000, *Arabidopsis thaliana* est une plante modèle pour la communauté internationale. Les raisons sont diverses. D'un point de vue expérimental, sa petite taille (entre 15 et 20 cm pour une plante adulte), son cycle de vie très court (6 semaines de graines à graines), sa grande résistance et l'aptitude des cellules à s'autoféconder facilitent sa culture et son étude

en laboratoire. D'un point de vue génétique, elle est un représentant des plantes car son ADN a été entièrement séquencé (il s'agit du premier génome végétal séquencé) et son génome est l'un des plus petits parmi ceux connus chez les végétaux (composé de 157 millions de paires de base réparties sur cinq paires de chromosomes). Enfin, d'un point de vue économique, elle ne représente aucun intérêt financier ce qui facilite le partage des connaissances à son sujet.

Figure 1.1: *Arabidopsis thaliana*

Si l'annotation structurale d'*Arabidopsis thaliana*, c'est-à-dire la description de l'ADN (sa séquence, la localisation des gènes,...) est à présent bien stable, il en est tout autre pour l'annotation fonctionnelle : seuls 16% des gènes ont une fonction validée expérimentalement et il reste encore 20% de gènes sans aucune information fonctionnelle [Zaag et al., 2015]. Quant à l'annotation relationnelle, qui consiste à identifier les interactions entre gènes mais aussi l'environnement dans lequel ces interactions se produisent, les connaissances sont parcellaires.

Toutefois, depuis une vingtaine d'année, les technologies à haut-débit sont apparues. Elles permettent de mesurer l'expression de tous les gènes simultanément et permettent de faire un changement d'échelle dans les analyses : car au lieu de se concentrer sur un petit ensemble de gènes, il est désormais maintenant possible de regarder le génome dans sa globalité.

Dans le cadre de cette thèse, nous avons cherché à identifier des liens de régulations entre certains gènes d'*Arabidopsis thaliana*, via une analyse de données transcriptomiques.

1.1.1.1 L'ADN : support de l'information génétique

Les cellules d'un organisme contiennent toutes, par définition, de l'acide désoxyribonucléique (ADN). L'ADN se situe dans le noyau de la cellule eucaryote. Cette macromolécule est constituée de deux brins complémentaires qui forment une double hélice. Chacun des brins constitue une séquence de *nucléotides* (A, T, C ou G) dont l'ordre est déterminant. L'ensemble de l'ADN forme le génome et certaines portions de cette séquence sont plus importantes : ce sont les régions codantes appelées aussi gènes ; les autres étant des régions non-codantes, appelées *régions intergéniques*.

L'information génétique est portée par la séquence complète de l'ADN. Celle-ci contient les caractères communs mais aussi les variabilités spécifiques de chaque individu. Pour le développement de ce dernier, son fonctionnement, sa reproduction et sa survie, l'information génétique doit s'exprimer. Pour cela, elle est transmise hors du noyau sous forme de protéines via l'expression des gènes.

L'expression d'un gène comporte deux étapes : la *transcription* et la *traduction*. Lors de la *transcription*, l'information génétique est transmise par recopiage de la séquence du gène en une séquence appelée Acides Ribonucléiques messager (ARNm) ou *transcrit*. Cet ARNm, formé des *nucléotides* U, A, G et C, sort du noyau pour atteindre le *cytoplasme* de la cellule. Lors de la *traduction*, l'ARNm est pris en charge par un ribosome qui lit la séquence de *nucléotides* et crée une protéine constituée d'acides aminés. L'ordre des *nucléotides* sur l'ADN détermine la succession des *acides aminés*. Un schéma récapitulatif de l'expression de gène est proposé Figure1.2.

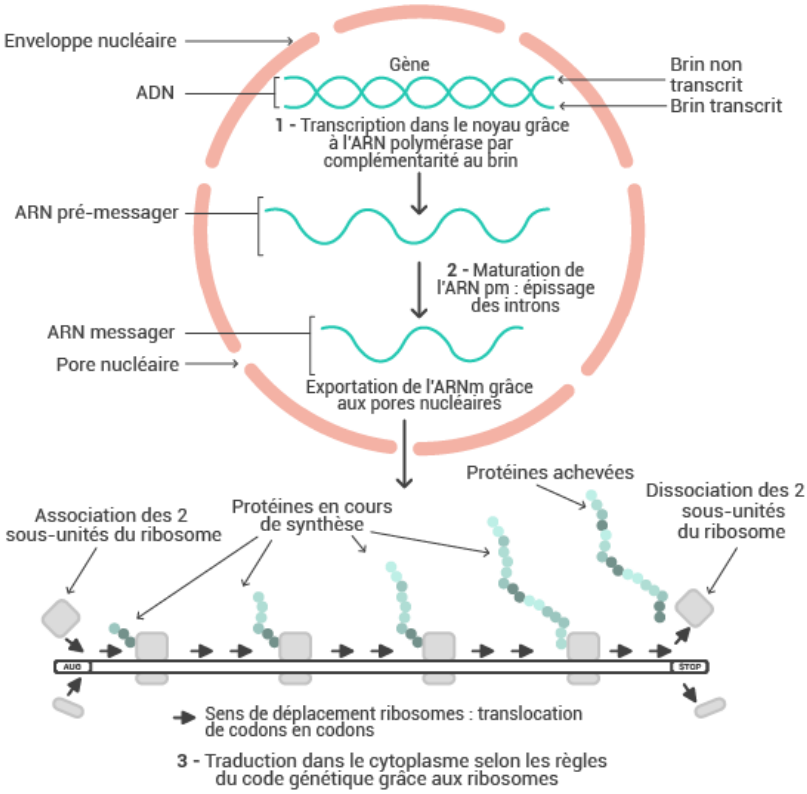


Figure 1.2: Mécanisme d'expression de gène [1]

L'expression des gènes est régulée pour créer uniquement les protéines nécessaires à un instant donné. Dans cette thèse, nous nous focalisons sur l'étude d'un mécanisme de régulation réalisé à l'échelle transcriptionnelle. Ce dernier implique les protéines *facteurs de transcription* (FTs) qui sont des composantes-clés de l'étape de *transcription* : elles permettent d'identifier la séquence

[1] `https://www.afterclasse.fr/fiche/269/lexpression-du-patrimoine-genetique/schema`

à coder et de démarrer la machinerie transcriptionnelle qui ouvre la double hélice de la séquence d'ADN, se fixe sur la séquence à exprimer, progresse le long de l'ADN, puis se détache. Ces protéines sont elles-mêmes issues de gènes. Par la suite, le terme *facteur de transcription* (FT) désignera, par abus de langage, le gène dont est issue la protéine FT.

1.1.1.2 Les *facteurs de transcription*

Un FT est donc un gène particulier. Une fois celui-ci *transcrit* puis traduit, sa protéine réintègre le noyau de la cellule pour se fixer sur la *région promotrice* (il s'agit de la séquence d'ADN qui précède celle du gène) d'un gène, appelé gène cible du FT. Les FTs sont connus pour travailler en module. Ainsi, un certain nombre de protéines sont nécessaires sur la *région promotrice* du gène cible pour démarrer sa transcription. Ce groupe de protéines forme un *complexe protéique*.

Du point de vue du gène cible, la présence de la protéine est essentielle pour démarrer sa *transcription*. Au sein du *complexe protéique*, certaines protéines sont *activatrices* et favorisent la *transcription* du gène cible, tandis que d'autres sont *inhibitrices* et empêchent la *transcription* (voir Figure1.3). Pour que la *transcription* démarre, le *complexe protéique* doit être complet au sens où une protéine de chaque FT du gène cible doit être présente. En d'autres termes, pour que la *transcription* démarre, il faut que tous les FT du gène cible s'expriment afin que leurs protéines soient fixées sur la *région promotrice* du gène cible. L'activité transcriptionnelle via les FTs est donc un acteur principal dans le mécanisme de régulation de l'expression génétique et est essentielle à la survie de la cellule.

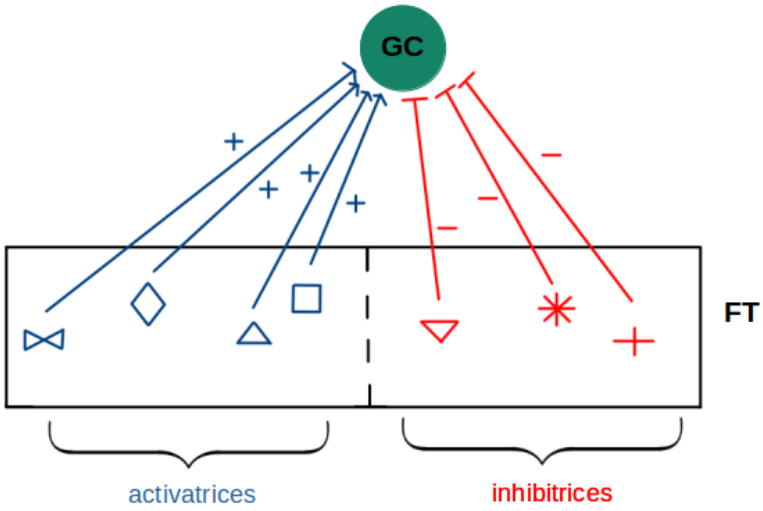


Figure 1.3: *Complexe protéique* présent sur la *région promotrice* d'un gène cible

1.1.1.3 Hypothèse biologique

Pour étudier la régulation des gènes, extraire de l'information sur la fixation des protéines issues des FTs semble une bonne idée. Pour cela, des technologies haut-débits existent. Par

exemple, les *expériences d'immunoprécipitation de chromatines (ChIP-seq)* [Barski et al., 2007] consistent à isoler une protéine pour localiser ses zones de fixation sur la séquence d'ADN. Si la protéine étudiée est issue d'un FT, alors les gènes cibles de celui-ci peuvent être identifiés. Une autre technique appelée *DNA affinity purification sequencing (DAP-seq)* [O'Malley et al., 2016] permet de récupérer les séquences d'ARNm où s'est fixée la protéine, ces séquences correspondent aux sites de fixation du FT correspondant. Malheureusement, ces techniques sont longues à réaliser et coûteuses. Par exemple, les expériences ChIP-seq exigent un *anticorps* spécifique à chaque protéine, et développer un *anticorps* n'est pas simple expérimentalement.

Dans cette thèse, nous apportons un nouveau regard sur la régulation des gènes par l'activité des FTs. Pour cela, nous nous tournons vers une approche complémentaire basée sur la mise en place de méthodes statistiques pour proposer des candidats FTs pour un gène cible donné. Cette approche prend en compte que les FTs travaillent en module ce qui a pour conséquence qu'un gène cible a plusieurs FTs, que chaque FT peut avoir différentes cibles et qu'un FT est aussi *transcrit* donc est lui-même cible de plusieurs FTs. Ainsi, de fortes dépendances existent donc entre les gènes et il est donc plus approprié d'étudier tous les FTs simultanément. Or, les approches expérimentales moléculaires ChIP-seq ou DAP-seq étudient les FTs un à un, ce qui ne permet pas l'étude globale de l'activité des FTs. De plus, nous souhaitons retrouver les FTs d'un gène cible. Or, les approches ChIP-seq et DAP-seq permettent de retrouver les cibles d'un FT et donc trouver les FTs pour un gène cible demande l'étude de tous les FTs un à un.

Il a déjà été montré que le transcriptome est une potentielle ressource d'information pour l'étude de la régulation [Vasseur, 2017, Mitsuda and Ohme-Takagi, 2009]. L'avènement de nouvelles technologies, comme le RNA-seq [Morin et al., 2008, Chu and Corey, 2012] ou la méthode des *puces à ADN* [Lenoir and Giannella, 2006], permet d'extraire de nouvelles données et d'acquérir des données transcriptomiques, soit une mesure représentative de la quantité de transcrits de tous les gènes simultanément. Nous gardons en tête que pour comprendre les liens de régulations entre les gènes, ce sont les sites de fixation des protéines issues des FT que nous devrions idéalement étudier. De plus, il n'y a malheureusement pas de liens systématiques entre la quantité d'un *transcrit* et celle de la protéine correspondante. Cependant, les données transcriptomiques sont acquises simultanément sur tous les gènes permettant l'étude globale de l'activité de ces derniers, grâce à l'application de procédures statistiques. C'est pourquoi, nous privilégions ces données dans cette thèse pour étudier le mécanisme de régulation de l'expression des gènes via les FTs en supposant que la quantité de *transcrits* mesurée est représentation de la quantité de protéines produites et donc du niveau d'expression du gène.

1.1.1.4 L'objectif biologique et les données disponibles

L'objectif biologique de cette thèse est de savoir si à partir des données transcriptomiques et d'un modèle statistique bien choisi, il est possible de retrouver les liens de régulation déjà connus chez *Arabidopsis thaliana* et d'éventuellement en identifier des nouveaux.

Pour cela, nous disposons des données publiques produites par la plateforme transcriptomique

POPS [2]. Ces données ont été pré-traitées en amont de cette thèse. Plus précisément, le jeu de données contient 2215 mesures d'expression indépendantes de 19844 gènes d'*Arabidopsis thaliana* dont 1935 sont des FTs. Ces mesures ont été prélevées sur 13 organes différents (racine, fleur, feuille, tige,...) et dans des conditions très variées. Par exemple, certaines conditions entraînent des mutations ou des stress biotiques (déclenchés par un être vivant : champignon, insecte, bactérie,...) ou abiotiques (induits par tout ce qui est non-vivant : la chaleur, le gel, la sécheresse, l'humidité,...). Les 2215 mesures d'expression indépendantes sont plus précisément des log-ratio entre l'expression des gènes dans une condition étudiée (stress biotique, abiotique, mutation) et l'expression des gènes dans un condition de référence.

Dans cette thèse, nous nous intéressons à 4 gènes cibles : LEAFY (AT5G61850), AP1 (AT1G69120), AP3 (AT3G54340) et AG (AT4G18960), qui sont eux-mêmes des FTs. L'objectif est d'identifier le complexe protéique formée sur leur *région promotrice* lors de la transcription. Ces gènes sont biologiquement connus pour interagir physiquement et pour jouer un rôle essentiel dans le développement floral d'*Arabidopsis thaliana* La nouveauté dans l'objectif biologique de cette thèse est de proposer, via des procédures statistiques appliquées sur les données transcriptionnelles, une liste de FTs pour chacun des 4 gènes cibles, de savoir si cette liste coïncide avec les liens de régulations déjà connus et si de nouveaux liens de régulation peuvent être proposés.

1.1.2 La régression linaire gaussienne en grande dimension

La problématique biologique est l'élément déclencheur de cette thèse : le contexte biologique motive l'ensemble des problématiques statistiques considérées. Naturellement, vue la quantité de données produites ainsi que leurs grandes variabilités dues à la fois aux appareils de mesure mais également à la variabilité intrinsèque du génome, la modélisation statistique est incontournable.

1.1.2.1 Les régressions linéaires gaussiennes

- La distribution gaussienne :
Le jeu de données disponibles contient $n = 2215$ réplicats biologiques sous des conditions expérimentales différentes. Plus précisément, ce sont des log-ratios de la mesure simultanée des quantités de *transcrits* du génome dans une condition (stress, mutations,...) sur la mesure simultanée des quantités de *transcrits* du génome dans une condition de référence. Comme le jeu de données ne contient pas d'autres réplicats techniques, nous disposons d'un échantillon de taille $n = 2215$ de données $\Big((y_1, (x_{1,1}, \cdots, x_{1,p}), \cdots, (y_n, (x_{n,1}, \cdots, x_{n,p})\Big)$ indépendantes. Nous modélisons par $Y \in \mathbb{R}^n$, le vecteur des données d'expression du gène cible et par $X \in \mathbb{R}^{np}$ la matrice constituée des n données d'expression des p autres FTs. Si Y est lui-même un FT (ce qui est le cas pour LEAFY, AP1, AP3 ou AG), alors X n'est constituée que des $p - 1$ FTs

[2] `https://ips2.u-psud.fr/fr/plateformes/spomics-interatome-metabolome-transcriptome/pops-plateforme-transcriptomique.html`

restants. Par soucis de simplification des notations, nous considérons par la suite que X est toujours de taille $n \times p$.

La Figure1.4 est un exemple d'histogramme obtenu sur les n données disponibles pour un gène. Cette distribution unimodale, symétrique et à queues légères justifie le choix de la distribution gaussienne pour la variable Y modélisant le gène cible, les variables $X_1, \cdots, X_p$ étant considérées fixes dans ce modèle. Utiliser la distribution gaussienne pour de telles données d'expression de gènes n'est pas nouveau [Wu and Ye, 2006, Vasseur, 2017, Yang et al., 2017].

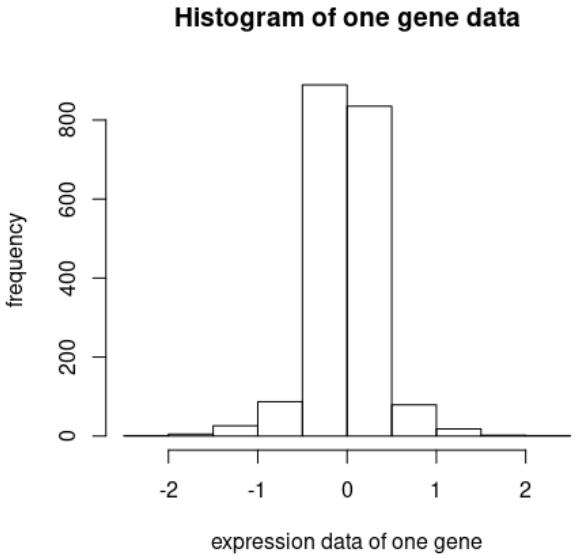


Figure 1.4: Un exemple de l'histogramme des n données disponibles pour un gène

- La corrélation de Pearson :
Les méthodes les plus populaires pour identifier des liens entre des gènes via des données transcriptomiques sont généralement construites sur l'utilisation de la corrélation de Pearson [Maertens et al., 2018]. Les corrélations simples paire à paire les plus élevées obtenues sur les données transcriptomiques donnent un ensemble de candidats FTs au gène cible étudié. Cependant, ces approches ne sont pas pertinentes puisque une forte corrélation de Pearson entre deux entités n'implique pas un lien direct : ce lien peut être indirect car il fait intervenir d'autres entités : si a agit sur b qui lui même agit sur c, alors a n'agit pas directement sur c, pourtant la corrélation de Pearson entre a et c a de fortes chances d'être élevée puisque la variation de a aura une conséquence sur celle de c. De même, si a agit sur b et c, la corrélation de Pearson entre b et c a de fortes chances d'être élevée, pourtant un lien direct n'est pas forcément existant.

- La corrélation partielle :
Pour identifier uniquement les liens directs, la corrélation partielle est adaptée. Elle a déjà été

longuement utilisée pour reconstruire le réseau de régulation entier d'un organisme [De La Fuente et al., 2004] [Peng et al., 2009]. Pour un lien entre a et b, le principe est de retirer tous les effets des autres variables sur a et sur b et de regarder s'il reste de la corrélation de Pearson entre les deux résidus. Si c'est le cas, alors de l'information commune entre a et b existe et elle leur est spécifique car non expliquée par les autres variables : dans ce cas, a et b sont dites partiellement corrélées. Si ce n'est pas le cas, alors il n'y a pas de lien direct entre a et b et les variables sont dites non partiellement corrélées.

- La régression linéaire gaussienne :
Lorsque les données sont gaussiennes, trouver les corrélations partielles entre Y et l'ensemble $(X_1, \cdots, X_p)$ est équivalent à trouver les variables impliquées dans la régression linéaire de Y contre $(X_1, \cdots, X_p)$ à un bruit gaussien centré près. Y est donc la variable régressée, aussi appelée variable réponse ; les variables $(X_1, \cdots, X_p)$ sont les régresseurs, aussi appelées variables explicatives.

Ainsi, le modèle statistique naturel, et celui utilisé dans cette thèse, est le suivant :

$$Y = X\beta^* + \varepsilon \tag{1.1}$$

où

- $Y \in \mathbb{R}^n$ et $X \in \mathbb{R}^{np}$ sont les variables de données.
- $\varepsilon \in \mathbb{R}^n$ est un vecteur aléatoire de bruit suivant la distribution $\mathcal{N}(0, \sigma^2 I_n)$, où I_n désigne la matrice identité de $\mathcal{M}_n(\mathbb{R})$.
- $\beta^* \in \mathbb{R}^p$ est le paramètre inconnu intervenant dans la régression.

On a donc $Y \sim \mathcal{N}(X\beta^*, \sigma^2 I_n)$.

A partir du n-échantillon $\Big((y_1, (x_{1,1}, \cdots, x_{1,p}), \cdots, (y_n, (x_{n,1}, \cdots, x_{n,p})\Big)$, trois problèmes statistiques différents existent sur la régression linéaire gaussienne : le problème d'estimation reposant sur l'estimation de β^*, le problème de prédiction reposant sur l'estimation de $X\beta^*$ et le problème de sélection reposant sur l'estimation du support de β^*, c'est-à-dire localiser les coordonnées non-nulles de β^*. De manière équivalente, ce dernier problème revient à retrouver les variables réellement impliquées pour expliquer Y. En effet, si $\beta_j^* \neq 0$, alors X_j intervient dans la relation linéaire, tandis que si $\beta_j^* = 0$, X_j n'y intervient pas. Ainsi, l'inférence du support de β^* revient à retrouver les corrélations partielles existantes entre Y et les variables $(X_1, \cdots, X_p)$. Obtenir β^* par le problème d'estimation permet, en étudiant le signe des coordonnées, de déterminer si la corrélation partielle est positive ($\beta_j^* > 0$) ou négative ($\beta_j^* < 0$) et ainsi de proposer des candidats de FTs activateurs pour le gène cible et des candidats de FTs inhibateurs pour le gène cible. Concernant le problème de prédiction, il est utilisé pour prédire y_{n+1} lorsque de nouvelles données $(x_{n+1,1}, \cdots, x_{n+1,p})$ sont disponibles.

- Les limites du modèle linéaire gaussian :
Bien que le modèle linéaire gaussian soit le modèle le plus simple en régression, celui-ci ne répond pas à l'objectif biologique. En effet, $\beta_j^* \neq 0$ correspond à la présence d'une interaction directe entre Y et X_j et traduit biologiquement un lien de régulation entre le gène cible et le FT numéro j. Cependant, comme la corrélation partielle est une quantité symétrique, le lien de régulation ne peut pas être orienté. Ainsi, si Y est lui même un FT, il est impossible de savoir lequel des deux régule l'autre. Pour gagner cette information supplémentaire, d'autres stratégies sont à déployer comme par exemple l'utilisation de données d'intervention (knock-out) ou de données temporelles via des méthodes bayésiennes par exemple [Chen et al., 2006]. De plus, les valeurs des coefficients non-nuls de β^* n'ont pas d'interprétation biologique : une valeur absolue élevée n'est pas liée à un lien de régulation fort. Ainsi, ce modèle ne permet que d'identifier la présence ou l'absence d'un lien de régulation entre un couple de deux gènes.

Pour la suite, nous définissons $\mathcal{D} = (Y, X)$ le jeu de données disponibles et
$m^* = \text{Vect}(X_j,\ j \text{ tel que } \beta_j^* \neq 0)$ le sous-espace vectoriel de $\mathbb{R}^p$ engendré par les variables réellement impliquées dans la relation linéaire pour expliquer Y.

1.1.2.2 Les défis de la grande dimension

Grâce aux progrès technologiques de ces dernières décennies, des bases de données gigantesques ont vu le jour et le nombre d'observations disponibles est sans cesse en train de croître. Parallèlement, des mesures sont collectées sur de plus en plus de variables et ce pour un même problème. Ainsi, le nombre de variables à prendre en compte dans un modèle statistique peut maintenant atteindre des dizaines voire des centaines de milliers. Par exemple, dans le cadre génomique, les nouvelles technologies permettent l'acquisition simultanée de données sur le génome complet d'un individu. Pour donner un ordre de grandeur, celui-ci contient 26000 gènes chez l'homme et 54000 gènes pour le maïs. Cependant, à cause du coût des expériences nécessaires pour extraire les données, le nombre d'observations reste très faible par rapport au nombre de variables disponibles nécessitant de plonger le modèle statistique dans un contexte de grande dimension. Depuis le début des années 2000, la statistique a cherché à s'affranchir des procédures classiques d'estimation pour s'adapter à ce nouveau paradigme ainsi qu'à la structure des données de plus en plus complexe. La sous-section 2.1 du chapitre 2 définit la notion de grande dimension et présente les difficultés engendrées par ce contexte.

Dans cette thèse, nous nous plaçons dans le cadre de la grande dimension : $p = 1887$, le nombre de variables candidates, est grand et proche de $n = 2215$. Cependant, nous prenons soin de vérifier que nous ne tombons pas dans le cadre de la *ultra haute-dimension* où les estimations sont impossibles à évaluer. N.Verzelen [Verzelen et al., 2012] a montré que ce cadre concerne tous les sous-ensembles de variables de taille $k > k^*$ où

$$k^* = \min\{k \in \mathbb{N}, 2k \log\left(\frac{p}{k}\right) \geq n\}$$

Ainsi, tout ensemble de taille supérieure à k^* est à retirer de la procédure statistique car les estimations associées sont mauvaises.

1.1.2.3 La sélection de variables

- L'estimateur des moindres carrés :
Notons $||.||_2$ la norme euclidienne usuelle sur $\mathbb{R}^n$. L'estimateur classique du paramètres β^* dans notre cadre de la régression linéaire gaussienne (1.1) est l'estimateur des moindres carrés (équivalent à l'estimateur du maximum de vraisemblance car la distribution est pour nous gaussienne) :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} ||Y - X\beta^*||_2^2, \tag{1.2}$$

et sa forme explicite est :

$$\hat{\beta} = (x^T x)^{-1} x^T y.$$

Lorsque la variance σ^2 est inconnue, son estimateur empirique dépend de $\hat{\beta}$ et vaut :

$$\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}||_2^2}{n-p}$$

Comme $\hat{\beta}$ et $\frac{n-p}{n}\hat{\sigma}^2$ sont les estimateurs du maximum de vraisemblance, on connaît leur loi : $\hat{\beta} \sim \mathcal{N}(\beta^*, \sigma^2(x^T x)^{-1})$ et $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$. De plus, par le *théorème de Cochran*, $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants. Cependant, l'hypothèse d'inversibilité sur $x^T x$ est nécessaire et celle-ci fait défaut systématiquement lorsque $p > n$; n'est pas toujours vérifiée dès que $p \approx n$ avec n grand, et lorsqu'elle l'est, son inversion demande un coût computationnel important. En grande dimension, séparer le signal β^* du bruit $\mathcal{N}(0, \sigma^2)$ est en général impossible. De plus, les propriétés de l'estimateur $\hat{\beta}$ (1.2) sont valables lorsque p est significativement plus petit que n et lorsque n tend vers l'infini : il est sans biais, consistant et asymptotiquement normal. Ainsi, ce contexte de grande dimension oblige à revoir la procédure d'estimation du vecteur des paramètres.

- La parcimonie :
Pour prendre en compte la grande dimension, de la régularité doit être imposée à l'estimateur. Souvent, la diminution significative de la dimension de l'espace considéré est nécessaire. Un exemple classique est l'*Analyse en Composantes Principales (ACP)* qui utilise des distances géométriques sur le nuage de points disponible [Saporta, 2006]. La sortie de cette approche est un ensemble de nouvelles variables qui sont des combinaisons linéaires des variables corrélées d'origine. Elles sont en petit nombre et résument l'information principale pour le problème statistique. Une alternative est de résumer l'information en utilisant, non pas des nouvelles variables, mais certaines parmi les variables candidates $(X_1, \cdots, X_p)$. Cette procédure est adaptée à notre question biologique : pour un gène cible donné et modélisé par Y, sa régulation est résumée par quelques variables parmi $(X_1, \cdots, X_p)$ qui correspondent à un set de FTs candidats. Cela revient à ajouter l'hypothèse de parcimonie au modèle de la régression linéaire gaussienne (1.1) : parmi toutes les variables candidates $(X_1, \cdots, X_p)$, peu d'entre elles

sont réellement impliquées pour expliquer Y et toutes les autres (qui représentent une majorité) sont soit inutiles car redondantes par rapport aux variables réellement impliquées dans la régression ; soit néfastes et dans ce cas, les considérer dans le modèle peut engendrer des erreurs d'estimation. Par la suite, une variable réellement impliquée pour expliquer Y au sens de la régression linéaire gaussienne sera dite active ; toutes les autres seront dites non actives. Pour le modèle statistique, cette hypothèse de parcimonie signifie que la taille du support de β^* est petit devant p (de nombreux coefficients sont nuls) : une variable active est indexée par un j tel que $\beta_j^* \neq 0$ tandis qu'une variable non active correspond à un $\beta_j^* = 0$. Ainsi l'estimation de β^* est réalisée parmi les vecteurs $\beta \in \mathbb{R}^p$ parcimonieux. D'après [Verzelen et al., 2012], si $|\beta^*| \leq \min\{k \in \mathbb{N}, 2k\log(\frac{p}{k}) \geq n\}$, alors l'estimation est possible. Tout l'enjeu est de localiser les coefficients non nuls de β^*, c'est-à-dire de retrouver l'ensemble des variables actives.

D'un point de vue pratique, cette hypothèse de parcimonie n'est pas aberrante. En effet, l'information pertinente sur les données est, dans de nombreux cas, contenue dans un espace de dimension beaucoup plus petite et est concentrée sur quelques variables seulement. Dans notre contexte biologique (Section 1.1.1), cela revient à supposer que le nombre de facteurs de transcription impliqués dans dans le processus de régulation d'un gène cible est très petit devant le nombre total de facteurs de transcription de la plante. Cette hypothèse est cohérente avec la connaissance biologique sur le processus de régulation des gènes via les facteurs de transcription. Finalement, l'ensemble des FTs candidats correspond à l'ensemble des variables déclarées actives lors de la procédure statistique.

- La sélection de variables :
Imposer la parcimonie de β^* incite à ré-écrire le problème d'estimation du paramètre comme un problème de sélection de variables. L'objectif est la mise en place d'une procédure statistique algorithmiquement faisable permettant d'identifier les variables actives. Dans la littérature ont été proposées deux approches différentes : l'approche prédictive via une procédure d'estimation, et l'approche FDR via une procédure de tests. Elles sont traitées indépendemment pour répondre à des problèmes différents.

1.2 Quelques éléments historiques sur la sélection de variables

Dans un contexte de grande dimension, la statistique fournit une réponse à la problématique biologique de cette thèse. Une procédure de sélection de variables est appliquée sur les données d'expression de gènes pour identifier une liste de candidats FTs pour un gène cible. Pour cela, deux approches ont été proposées en parallèle : l'approche prédictive reposant sur l'estimation du vecteur $X\beta^*$ et l'approche par les tests multiples reposant sur l'estimation du support de β^*. Ces deux approches répondent à des problèmes différents et minimisent, pour l'un le *risque prédictif (RP)* (sous-section 1.2.1), et pour l'autre le *False Discovery Rate (FDR)* (sous-section 1.2.2). Des procédures statistiques ont également vu le jour pour tenter de combiner ces deux approches (sous-section 1.2.3).

1.2.1 L'approche prédictive par la sélection de modèles

L'approche naturelle du modèle de régression linéaire gaussienne dans un contexte de grande dimension est de s'inspirer de la minimisation classique des moindres carrés en y ajoutant la contrainte de sélection de variables. Le contraste choisit, les moindres carrés, dont la minimisation équivaut à la maximisation de la log-vraisemblance du modèle dans le cadre gaussien, évalue la qualité de l'ajustement linéaire de $\hat{\beta}$ pour expliquer Y. Pour cette approche, la qualité de l'estimation finale de $\hat{\beta}$ est évaluée par le risque prédictif :

$$\mathbb{E}_Y[||X\hat{\beta} - X\beta^*||_2^2].$$

La prédiction est donc ce qu'il y a de plus naturel en grande dimension : l'estimateur final $\hat{\beta}$, construit sur des observations $(y_1, \cdots, y_n, x_{1,1}, \cdots, x_{1,p}, \cdots, x_{n,1}, \cdots, x_{n,p})$, doit permettre la prédiction correcte d'une nouvelle valeur y_{n+1} à partir de nouvelles observations associées $\big(x_{n+1,1}, \cdots, x_{n+1,p}\big)$.

- La sélection de modèles :
Pour sélectionner le meilleur groupe de variables d'un point de vue prédictif, les premières approches, proposées dans les années 70, consistent à sélectionner le meilleur sous-ensemble parmi une collection de sous-ensembles donnée, notée $\mathcal{M}$. Un modèle $m \in \mathcal{M}$ est défini comme le sous-espace vectoriel de $\mathbb{R}^p$ engendré par certaines variables $(X_j)_{j\in\{1,\cdots,p\}}$. La dimension de m est notée D_m et l'estimateur des moindres carrés à l'intérieur de m est noté $\hat{\beta}_m$ et vaut

$$\hat{\beta}_m = \underset{\{\beta, X\beta\in m\}}{\arg\min} ||Y - X\beta||_2^2.$$

Le modèle sélectionné $\hat{m}$ est celui qui réalise le meilleur compromis entre l'ajustement linéaire de $X\hat{\beta}_m$ pour expliquer Y et la parcimonie de $\hat{\beta}_m$. Pour l'obtenir, $\hat{m}$ est le modèle qui résout le problème d'optimisation suivant : pour $\mu > 0$,

$$\hat{m} = \underset{m\in\mathcal{M}}{\arg\min} \Big\{ ||Y - X\hat{\beta}_m||_2^2 + \mu D_m \Big\}. \tag{1.3}$$

Plus μ est grand, plus la parcimonie est forte et moins l'ajustement linéaire est bon ; plus μ est petit, plus la parcimonie est faible et meilleur est l'ajustement linéaire. Historiquement, [Akaike, 1973] et [Schwarz et al., 1978] sont les premiers à avoir proposé des valeurs pour μ dans un cadre plus général :

$$\mu_{\text{AIC}} = 2\sigma^2$$

pour *AIC* [Akaike, 1973] et

$$\mu_{\text{BIC}} = \log(n)\sigma^2$$

pour *BIC* [Schwarz et al., 1978]. Au même moment, [Mallows, 2000] étudie cette approche dans le cadre de la régression et obtient la même valeur qu'AIC :

$$\mu_{\text{CP\_Mallows}} = 2\sigma^2$$

pour *CP.Mallows*. Dès que $\mathcal{M}$ devient trop grande (par exemple, si $\mathcal{M}$ contient plus d'un modèle par dimension), ces pénalités ne sont pas assez fortes et $\hat{m}$ a tendance à être trop grand : l'estimation tombe dans la zone de sur-apprentissage. Ce phénomène a par exemple été observé biologiquement avec le critère BIC (AIC et CP.Mallows étant encore plus libéraux) lors de l'identification de locus génétiques (*Quantitative Trait Loci (QTL)*) contribuant à la variation d'un trait quantitatif dans des croisements expérimentaux [Broman and Speed, 2002, Bogdan et al., 2004]. De plus, les propriétés de $\hat{m}$ sont asymptotiques : elles garantissent que μ_{AIC} et $\mu_{\text{CP\_Mallows}}$ offrent le plus petit risque prédictif asymptotiquement, et que μ_{BIC} offre la consistance de l'estimateur. Dans une toute autre direction, [Birgé and Massart, 2001] se tourne vers une forme plus générale sur la contrainte de parcimonie et considèrent le problème d'optimisation suivant : trouver $\hat{m}$ tel que :

$$\hat{m} = \underset{m \in \mathcal{M}}{\arg\min} \left\{ ||Y - X\hat{\beta}_m||_2^2 + \text{pen}(D_m) \right\} \tag{1.4}$$

où pen est une fonction positive et croissante en D_m. Ils proposent ainsi une fonction pen_{min} tel que pour tout $\hat{m}$ solution de (1.4) obtenu avec pen telle que $\text{pen}(m) \geq \text{pen}_{\text{min}}(m)$ sur $\mathcal{M}$, une inégalité oracle non-asymptotique est vérifiée. Cela signifie que quel que soient n et p fixés, le risque prédictif associé à $\hat{m}$ est plus petit, à constantes près, que le plus petit risque prédictif qui peut être obtenu sur la collection de modèles. Cette pénalité pen_{min} est donnée sous une forme explicite et fait intervenir les propriétés de $\mathcal{M}$. En particulier, le terme $\log\left(\binom{p}{D_m}\right)$ est introduit dans la fonction de pénalité rendant compte du nombre de modèles disponibles par dimension. Ainsi, leurs travaux offrent une infinité de fonctions de pénalité vérifiant des performances prédictives pour tous n et p fixés. Par exemple, toute fonction $\text{pen} = K\text{pen}_{\text{min}}$ pour $K > 1$ garantie une inégalité oracle non-asymptotique sur le risque prédictif. Dans cette direction, [Chen and Chen, 2008] propose une nouvelle valeur de μ pour (1.3), extension de μ_{BIC}, faisant intervenir ce terme logarithmique pour atteindre la propriété de consistance asymptotique pour une grosse collection de modèles $\mathcal{M}$. Une de leurs applications motrices en biologie moléculaire est l'identification de *polymorphismes* d'un *nucléotide* (SNP) (parmi un nombre de candidats allant jusqu'à l'ordre du 10^7) responsables d'une variation à l'échelle génotypique [Marchini et al., 2005].

Si $\mathcal{M}$ est la collection de modèles complète, c'est-à-dire contenant tous les sous-ensembles de variables possibles parmi $(X_1, \cdots, X_p)$, alors le problème (1.3) est équivalent au suivant : pour un $\lambda > 0$ bien choisi,

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ ||Y - X\beta||_2^2 + \lambda |\beta|_0 \right\}, \tag{1.5}$$

où $|.|_0$ est la norme ℓ_0 usuelle de $\mathbb{R}^p$. Par exemple, minimiser le critère régularisé (1.5) avec $\lambda = 2\sigma^2$ est équivalent à minimiser (1.3) sur la collection de modèles complète avec $\mu = \mu_{\text{CP\_Mallows}}$. Cette réécriture montre que le critère qui est minimisé est non convexe et non dérivable, ceci à cause de l'irrégularité de la norme ℓ_0. Par conséquent, ces problèmes d'optimisation combinatoires sont NP-hard avec un temps de calcul qui augmente exponentiellement avec la dimension p [Natarajan, 1995]. Ainsi, dès que p est grand (typiquement quelques dizaines), voire dépasse n, le problème d'optimisation dans l'espace entier $\mathbb{R}^p$ est impossible à

résoudre en temps raisonnable et il est nécessaire d'imposer de la régularité au critère à minimiser pour obtenir des algorithmes à temps computationnel linéaire en la dimension.

- La sélection de modèle pour la grande dimension par régularisation :
Dans les années 90, [Tibshirani, 1996] propose de remplacer la norme ℓ_0 par la norme ℓ_1. Le problème de minimisation devient convexe et dérivable sauf en 0 et est appelé le *critère Lasso*. L'estimateur sous-jacent est $\hat{\beta}_{\lambda,\text{Lasso}}$ pour un λ donné. La norme ℓ_1 est un substitut efficace à la norme ℓ_0 puisque le Lasso sélectionne également des variables. En effet, pour un $\lambda > 0$ fixé, le Lasso estime à zéro les coefficients de $\hat{\beta}_{\lambda,\text{Lasso}}$ en dessous d'un certain seuil qui dépend de λ. Les variables correspondantes sont donc retirées de l'ensemble des variables sélectionnées. Trouver le $\lambda > 0$ qui réalise le meilleur compromis entre l'ajustement linéaire et la parcimonie est un réel défi. Une première manière est de fixer une valeur de λ pour respecter des propriétés théoriques. Dans le cadre de la régression gaussienne, si λ est choisi proportionnel à $\sigma\sqrt{\frac{\log(p)}{n}}$, alors des contrôles du risque prédictif sont obtenus sous différentes hypothèses [Bunea et al., 2007a,Bunea et al., 2007b]. Par exemple, si $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$ avec $A > 2\sqrt{2}$, alors une inégalité oracle sur $||X\beta^* - X\hat{\beta}_{\lambda,\text{Lasso}}||_2^2$ est vérifiée avec grande probabilité sous une condition de valeur propre restreinte [Bickel et al., 2009]. Un choix de λ proportionnel à $\sigma\sqrt{\frac{\log(p)}{n}}$ vérifie des propriétés similaires sur le risque prédictif dans d'autres modèles : voir [Bunea et al., 2006] pour une régression plus générale où le bruit n'est plus nécessairement gaussien, [Bertin et al., 2011] pour l'estimation de densité en régression, et [Ivanoff et al., 2016] dans le cadre de la régression fonctionnelle poissonnienne. Dans une autre direction, [Zou, 2006] propose le *Lasso adaptatif* en pondérant la pénalité ℓ_1. Les variables $(X_1, \cdots, X_p)$ n'ont ainsi pas la même importance lors de la minimisation Lasso. Un choix de λ adapté pour le Lasso adaptatif permet de satisfaire des propriétés asymptotiques et de consistance sous des conditions moins restrictives que le Lasso.

Cependant, malgré un bon choix de λ, l'estimateur $\hat{\beta}_{\lambda,\text{Lasso}}$ obtenu ne minimise pas les moindres carrés mais les moindres carrés régularisés. Le seuillage brutal des coefficients petits qui sont mis à zéro biaise l'estimateur $\hat{\beta}_{\lambda,\text{Lasso}}$. Pour pallier ce problème, une étape supplémentaire, mentionnée par [Efron et al., 2004] et explorée par [Connault, 2011], consiste à ré-estimer les coefficients non nuls $\hat{\beta}_{\lambda,\text{Lasso}}$ par minimisation des moindres carrés sur le support de $\hat{\beta}_{\lambda,\text{Lasso}}$. Le nouvel estimateur est alors asymptotiquement sans biais.

Une alternative au choix de λ consiste à rejouer la résolution du problème d'optimisation Lasso sur des réplicats du jeu de données initial. Cette approche s'inscrit dans la lignée des travaux qui utilisent le Lasso uniquement pour obtenir un support de variables. La *validation-croisée* [Refaeilzadeh et al., 2009] est l'approche la plus célèbre mais celle-ci est coûteuse et fournit des résultats pauvres en grande dimension. [Meinshausen and Bühlmann, 2010] d'une part et [Bach, 2008] de l'autre proposent des réplicats respectivement par sous-échantillonnages de taille $\lfloor \frac{n}{2} \rfloor$, obtenus par tirage uniforme et sans remise parmi le jeu de donnée d'origine de taille n (*Stability Selection*), et par ré-échantillonages de taille n, obtenus par tirage uniforme et avec remise (*Bolasso* pour BOotstrap-enhanced Least Absolute Shrinkage Operator). Leur motivation était de déjouer l'instabilité de l'approche Lasso puisqu'une légère modification des données peut

modifier drastiquement l'estimation de Lasso. Par leur approche, les variables sélectionnées correspondent aux variables les plus fréquemment sélectionnées lors de la résolution du Lasso sur chacun des réplicats. Les auteurs montrent que le choix de λ lors de la résolution du Lasso sur chacun des réplicats a peu d'importance sur l'estimation finale : quelque soit la valeur du λ pris dans un intervalle raisonnable (valeur ni trop grande, ni trop petite), les variables les plus fréquemment sélectionnées sont quasiment les mêmes. Ces variables sélectionnées sont associées à un support pour $\hat{\beta}$ dont les coefficients non nuls sont estimés par minimisation des moindres carrés sur ce support, dans la lignée de ce que proposent [Efron et al., 2004] et [Connault, 2011]. Une amélioration de la méthode Stability Selection a été proposée par [Haury et al., 2012]. Cette nouvelle méthode, appelée *Tigress* (pour Trustful Inference of Gene REgulation with Stability Selection), a été pensée pour inférer la structure d'un réseau de régulation de gènes. L'idée est d'appliquer le principe de Stability Selection en minimisant le critère Lasso sur chaque sous-échantillon et sur une grille bien définie de différentes valeurs de λ. Résoudre le Lasso sur une grille Λ de λ donne ce qui est couramment appelé le *chemin de régularisation* :

$$\left[\lambda \in \Lambda \mapsto \hat{\beta}_{\lambda,\text{Lasso}}\right].$$

Pour chaque variable, un score est calculé reflétant sa fréquence d'apparition dans chaque sous-échantillon mais aussi le long de chaque chemin de régularisation. Ainsi, une variable active est sélectionnée pour la plupart des sous-échantillons et à partir d'une grande valeur de λ jusqu'aux valeurs proches de 0. Au contraire, une variable non active n'est pas nécessairement sélectionnée avec tous les sous-échantillons et ne le sera que pour des valeurs petites de λ. Cette étape supplémentaire à Stability Selection permet à Tigress de gagner en robustesse.

En plus de la difficulté à trouver une bonne valeur de λ, la norme ℓ_1 a aussi été discutée. Ainsi, la norme ℓ_2 a été proposée par [Hoerl and Kennard, 1988] mais aussi la combinaison convexe des normes ℓ_1 et ℓ_2 : $(1-\alpha)|\beta|_1+\alpha||\beta||_2^2$ pour $\alpha \in [0,1]$ [Zou and Hastie, 2005]). Ici, $|.|_1$ et $||.||_2$ désignent respectivement les normes 1 et 2 usuelles sur $\mathbb{R}^p$. Cette combinaison convexe, qui donne le problème de minimisation appelé *Elastic-Net*, prend en compte la corrélation entre les variables puisqu'elle encourage la sélection par groupe de variables corrélées. L'introduction de cette nouvelle pénalité a été motivée par le même contexte biologique que celui considéré dans cette thèse : la sélection de gènes via des données de *puces à ADN*. Puisque les gènes ayant les mêmes voies biologiques peuvent partager de fortes corrélations entre eux [Segal et al., 2003], il est naturel de vouloir les sélectionner par groupe de gènes corrélés. Le Lasso n'est pas adapté pour cela car il ne sélectionne qu'au plus n variables sur les p candidats [Efron et al., 2004] et a tendance à ne sélectionner qu'une variable au hasard parmi un groupe de variables corrélées. Ainsi, Elastic-Net est préférable. Dans la même lignée, [Yuan and Lin, 2006] propose le Groupe Lasso où la parcimonie considérée est une parcimonie de groupe. Pour cette dernière approche, les groupes sont formés de variables corrélées mais ils doivent être connus a priori.

1.2.2 L'approche FDR issue des tests multiples

Dans cette thèse, nous étudions le False Discovery Rate (FDR) initialement introduit pour les tests multiples. Nous ne considérons pas de manière générale les tests multiples dans cette

thèse. C'est pourquoi, une étude historique des méthodes de tests multiples ne sera pas proposée (voir par exemple l'introduction de thèse de [Durand, 2018]) ou l'article de [Roquain, 2011]. L'objectif de cette sous-section se restreint à la genèse du FDR.

Notre jeu de données réel est constitué de n données d'expériences indépendantes d'un gène cible modélisé par Y et de p FTs candidats modélisés par les variables $(X_1, \cdots, X_p)$. Pour déterminer l'ensemble des variables actives dans la régression linéaire gaussienne, p tests sont à réaliser. Les *hypothèses nulles* $H_{0,j} = \{\beta_j^* = 0\}$ sont à tester contre les *hypothèses alternatives* $H_{1,j} = \{\beta_j^* \neq 0\}$ pour chaque $j \in \{1, \cdots, p\}$. Rejeter une hypothèse $H_{0,j}$ signifie que la variable X_j est déclarée active. Au contraire, conserver l'hypothèse $H_{0,j}$ signifie que le test n'a pas découvert X_j comme variable pertinente, elle est donc déclarée non active. Les variables aléatoires usuellement utilisées lors des procédures de tests sont les *p-valeurs* : $(\widehat{p}_1, \cdots, \widehat{p}_p)$. Pour chaque $j \in \{1, \cdots, p\}$, $\widehat{p}_j$ est la probabilité sous la distribution de $H_{0,j}$ d'obtenir une valeur au moins aussi extrême qu'une quantité, proprement définie, qui résume les observations du jeu de données disponible (dans notre contexte, les observations pour le test j sont les log-ratio des données d'expression du gène j et du gène cible). Ainsi, plus $\widehat{p}_j$ est petite, plus la quantité est extrême par rapport à la distribution sous $H_{0,j}$. Elle est donc peu probable sous $H_{0,j}$: le test rejette $H_{0,j}$ et accepte $H_{1,j}$ avec grande confiance. Au contraire, pour $\widehat{p}_j$ grande, alors l'hypothèse $H_{1,j}$ n'est pas assez vraisemblable et le test choisit de conserver $H_{0,j}$. Les hypothèses alternatives vraies sont donc associées avec forte probabilité aux plus petites p-valeurs. Ainsi, trancher entre les $H_{0,j}$ ou les $H_{1,j}$ revient à déterminer un seuil sur les p-valeurs en dessous duquel les $H_{0,j}$ sont rejetées et $H_{1,j}$ acceptées, et au dessus duquel les $H_{0,j}$ sont conservées.

La détermination du seuil est un réel défi, d'autant plus que lorsque p tests sont réalisés pour une réponse globale (dans notre cas, la réponse est une liste de FTs candidats parmi les p possibles), le critère d'erreur à contrôler doit lui aussi être global. Le critère naturel est le *Family Wise Error (FWER)*. Il s'agit de la probabilité qu'au moins une variable soit déclarée active à tort à l'issu des p tests. Si pour chaque j de 1 à p l'hypothèse $H_{1,j}$ est acceptée si et seulement si $\widehat{p}_j \leq \alpha$ (on dit que le test est de niveau α), alors le FWER n'est contrôlé que par $p\alpha$. Cette quantité est grande puisque p est grand. Ainsi, dans un contexte de grande dimension, les tests ne peuvent pas être considérés individuellement les uns des autres mais doivent être traités simultanément pour contrôler un critère global. Dans cette direction, [Bonferroni, 1936] propose dans les années 30 une correction, appelée correction de Bonferroni : ne sont rejetées que les $H_{0,j}$ associées à des $\widehat{p}_j < \frac{\alpha}{p}$. Dans ce cas, le FWER est contrôlée par α. Cette correction est équivalente à réaliser les p tests avec un niveau $\frac{\alpha}{p}$, quantité très petite pour p grand, ce qui rend la procédure très conservative. Des modifications à la correction de Bonferroni ont été proposées [Dunn, 1961, Holm, 1979, Simes, 1986, Hommel, 1988, Hochberg, 1988, Rom, 1990] mais ce qui a révolutionné le domaine de la biologie est le False Discovery Rate (FDR) introduit par [Benjamini and Hochberg, 1995]. Il est défini par :

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#\Big\{j,\ H_{0,j} \text{ rejetée à tort}\Big\}}{\#\Big\{j,\ H_{0,j} \text{ rejetée}\Big\} \vee 1}\right].$$

Contrôler le FDR permet d'augmenter le nombre de variables déclarées actives en s'autorisant une petite proportion de variables déclarées actives à tort parmi les variables déclarées actives. Dans leur article [Benjamini and Hochberg, 1995], les auteurs proposent un seuil sur les p-valeurs de sorte que le FDR soit contrôlé par α mais sous l'hypothèse que les p-valeurs sont des variables aléatoires indépendantes. Cette hypothèse revient à supposer l'absence de corrélation entre les variables $(X_1, \cdots, X_p)$. Des résultats sont connus sous des hypothèses plus souples mais les corrélations entre les variables restent très contrôlées (par exemple satisfaire la condition suffisante PRDS (Positive Regression Dependence on a Subset)) [Benjamini and Yekutieli, 2001, Romano et al., 2008].

Les procédures de tests multiples sont massivement utilisées en pratique et nous renvoyons aux livres [Dudoit and van der Laan, 2008], [Goeman and Solari, 2014] et [Cui et al., 2021] pour de nombreuses applications des procédures de tests multiples notamment en génomique et en médecine. Par exemple, pour déterminer si la différence d'expression de p gènes varie ou non entre deux conditions expérimentales A et B, p tests sur la moyenne des différences d'expression peuvent être réalisés simultanément et être facilement contrôlés via les procédures de tests multiples. Une autre application génomique où le FDR est massivement utilisé est l'*étude d'association pangénomique* (*Genome-Wide Association Studies (GWAS)*). Un *polymorphisme mononucléotidique* (*Single Nucleotide Polymorphism (SNP)*) est la variation d'un *nucléotide* unique au sein d'individus d'une même espèce et sur une région bien précise de l'ADN. Chez l'Homme, les SNPs sont à l'origine de la plupart des variations génétiques [Gibbs et al., 2003]. Une étude GWAS consiste à associer les SNPs à des phénotypes d'intérêt (une maladie par exemple). Pour cela, un test statistique est réalisé par SNP : l'hypothèse nulle correspond à une absence d'association entre le SNP et le trait phénotypique tandis que l'hypothèse alternative correspond à la présence d'une association. Considérer l'ensemble des tests simultanément est essentiel pour limiter le nombre de SNPs identifiés à tort. Les GWAS sont massivement étudiées [MacArthur et al., 2017] et sont une application pertinente des tests multiples puisque les SNPs sont environ 10^6, ce qui plonge le problème statistique dans un cadre de grande dimension ; et ils contiennent de la dépendance puisque deux SNPs proches sur l'ADN sont liés, ce qui empêche de considérer les tests individuellement les uns des autres.

1.2.3 Combiner risque prédictif et FDR

Généralement, les critères du risque prédictif et du FDR sont contrôlés seuls (sous-sections 1.2.1 et 1.2.2). D'ailleurs, [Yang, 2005] a soulevé l'impossibilité d'obtenir simultanément les performances prédictives asymptotiques parfaites et la consistance de l'estimateur. Mais, plutôt que de chercher à atteindre l'optimalité de ces deux propriétés, des récents travaux proposent d'obtenir un compromis entre les deux points de vue. Nous présentons ici une liste non-exhaustive dans notre cadre de la régression linéaire gaussienne en grande dimension.
Une première approche est une procédure multi-étape. [Zhou, 2009] propose de combiner la résolution du Lasso en choisissant $\lambda = \mathcal{O}\Big(\sigma\sqrt{2\frac{\log(p)}{n}}\Big)$ avec une procédure de seuillage sur les coefficients de l'estimateur obtenu. Une borne sur la perte ℓ_2 entre $\hat{\beta}$ et β^* est obtenue tout en

conservant les performances de prédiction du Lasso.
Une autre approche est l'inférence post-sélection. L'idée principale est de tester la pertinence de la sortie d'un algorithme appliqué sur les données. Dans notre cadre, le principe est de fournir des intervalles de confiance en post-sélection valides de niveau $1-\alpha$, pour un α donné, sur n'importe quel estimateur $\hat{\beta}_\lambda$ du chemin de régularisation (Lasso, Elastic-Net,...). Ceci via des tests d'hypothèses ou des tests d'hypothèses conditionnels (introduit dans [Pötscher, 1991]). Ces méthodes ont donné naissance aux intervalles de confiance connus dans la littérature sous le nom d'intervalles de confiance en inférence en post-sélection (*post-selection inference (PoSI)*) [Berk et al., 2013] et intervalles de confiance en inférence en post-sélection exacte (*exact post-selection inference (EPoSI)*) [Lee et al., 2016]. Cette approche a ensuite été généralisée par exemple en régression *forward-stepwise* dans [Tibshirani et al., 2016], en régression linéaire gausssienne généralisée dans [Hyun et al., 2018] et [Duy and Takeuchi, 2021], en modèle graphique gaussien dans [Chen et al., 2021]. Nous renvoyons le lecteur à [Zhang et al., 2022] pour une large étude des méthodes d'inférence post-sélection existantes sous le modèle de régression linéaire.
Dans une toute autre direction, [Abramovich et al., 2006] proposent une alternative aux pénalités régularisées comme le Lasso ou l'Elastic-Net en régression gaussienne multivariée de grande dimension à variance connue. Ils proposent une nouvelle forme de pénalité et une calibration de l'hyperparamètre en utilisant des outils des tests multiples, notamment les quantiles. L'estimation sous-jacente vérifie un contrôle du FDR et une égalité asymptotique minimax sur le risque prédictif.
En parallèle, les auteurs de [Bogdan et al., 2013] observent que pour un λ fixé, le Lasso s'apparente à la correction de Bonferroni en tests multiples où chaque p-valeur est comparée à une même valeur fixe. Ainsi, pour imiter l'amélioration proposée par [Benjamini and Hochberg, 1995] où les p-valeurs sont comparées à des seuils différents en fonction de leur rang quand elles sont triées dans l'ordre croissant, les auteurs proposent la pénalité $\lambda_1|\beta|_{(1)}+...+\lambda_p|\beta|_{(p)}$ comme alternative au Lasso. Ici, $\lambda_1 \geq ... \geq \lambda_p$ et $|\beta|_{(1)} \geq ... \geq |\beta|_{(p)}$ de telle sorte que plus la p-valeur associée à une variable X_j est grande, plus λ_j est grand, donc plus X_j est pénalisée et plus sa probabilité d'être sélectionnée est faible. Cette procédure est connue sous le nom de *SLOPE* (pour Sorted ℓ_1 penalized estimator). Sous le même modèle que [Abramovich et al., 2006] et lorsque les variables $\left(X_1, \cdots, X_p\right)$ sont orthogonales, l'estimateur SLOPE satisfait une égalité asymptotique minimax sur le risque prédictif et son FDR est contrôlé par α pour un bon choix du vecteur $\left(\lambda_1, ..., \lambda_p\right)$. Des résultats sont obtenus en remplaçant l'hypothèse d'orthogonalité par un contrôle très strict sur la corrélation entre les variables. Dans ce cas, les valeurs pour le vecteur $\left(\lambda_1, ..., \lambda_p\right)$ ne sont plus les mêmes. Ces résultats ont ensuite été étudiés dans des modèles plus généraux, par exemple quand la matrice de design X aléatoire [Kos and Bogdan, 2020].
Enfin, les auteurs de [Barber and Candès, 2015] adoptent une approche radicalement différente de la procédure FDR contrôlée par les tests multiples dans le cadre où p reste petit face à n. Ils proposent la *méthode des knockoffs* dont le principe est d'appliquer la procédure Lasso à la matrice augmentée $X = \left(X_1, \cdots, X_p, \tilde{X}_1, \cdots, \tilde{X}_p\right)$. Les $\tilde{X}_j$ sont des copies des X_j de telle

sorte que la structure de corrélation entre les copies soit identique à celle entre les variables d'origine mais que les $\tilde{X}_j$ soient complètement indépendantes de la variable réponse Y. L'idée clé est qu'une variable active apparaît très tôt dans le chemin de régularisation du Lasso alors que sa copie, indépendante de Y, apparaît bien plus tard. Quant aux variables non actives, elles apparaissent de manière équiprobable soit avant soit après leur copie. Une statistique de test est calculée sur chaque couple $(X_j, \tilde{X}_j)$ et elle tient compte de leur arrivée dans le chemin de régularisation. Un estimateur du FDR est proposé et est utilisé pour définir un seuillage sur les statistiques donnant l'ensemble des variables sélectionnées. Le FDR est contrôlé de manière exacte. Le principal avantage de la méthode des knockoffs est sa flexibilité puisqu'elle est construite sans connaissance sur la matrice de design X, sur l'amplitude du signal β^* et sur le niveau de bruit σ. Elle a ensuite été généralisée au cadre de la grande dimension [Barber and Candès, 2019, Gégout-Petit et al., 2019] ou au modèle non linéaire [Candès et al., 2016, Huang and Janson, 2020]. Des propriétés de robustesse ont également été prouvées [Barber et al., 2020].

1.2.4 Objectifs de la thèse

Cette thèse répond à trois objectifs.

Le premier objectif concerne uniquement la sélection de variables par l'approche prédictive via la sélection de modèles (comme décrit en sous-section 1.2.1). Nous proposons de calibrer l'hyperparamètre $\lambda > 0$ lors de la minimisation du critère régularisé en adaptant le point de vue de la sélection de modèles à la "Birgé-Massart" où toutes les fonctions pen proportionnelles à $\text{pen}_{\min}$ ($\text{pen} = K\text{pen}_{\min}$ pour $K > 1$) vérifient une inégalité oracle non-asymptotique sur le risque prédictif. Nous considérons ces deux approches dans l'ordre inversé par rapport à l'historique : la théorie de L.Birgé et P.Massart est appliquée après la résolution du problème d'optimisation de l'équation régularisée.
La combinaison de la théorie de régularisation et de la théorie de L.Birgé et P.Massart permet à cette dernière d'être appliquée dans un contexte de grande dimension où p peut être très grand (quelques dizaines) voire dépasser n et sur une combinatoire raisonnable. En effet, l'étape de régularisation permet de se restreindre à une collection ne contenant que quelques sous ensembles pertinents de variables, et non plus les 2^p modèles possible qui sont impossibles à explorer de manière exhaustive en toute généralité.
La combinaison de la théorie de L.Birgé et P.Massart et de la théorie de régularisation permet à cette dernière d'obtenir une calibration de l'hyperparamètre λ avec des garanties prédictives non-asymptotiques et en faisant intervenir une pénalisation ℓ_0, la norme idéale pour sélectionner des variables.

Le second objectif consiste à combiner la sélection de variables par l'approche prédictive via la sélection de modèles (sous-section 1.2.1) et le FDR introduit dans la sous-section 1.2.2. L'objectif est d'obtenir un compromis entre un contrôle du risque prédictif et un contrôle du FDR. Dans la lignée des travaux décrits en sous-section 1.2.3 et qui prennent en compte simultanément le contrôle du risque prédictif et le contrôle du FDR, nous proposons d'ajouter un

contrôle non-asymptotique du FDR dans la procédure de sélection de modèle tout en conservant ses performances de prédiction non-asymptotiques sur l'estimation.

Le troisième objectif de cette thèse est l'application sur données réelles de différentes méthodes statistiques étudiées dans cette thèse. Les FTs candidats proposés sont ensuite comparés à une liste de FTs connus biologiquement pour déterminer si les données transcriptomiques sont pertinentes pour comprendre les liens de régulation entre les gènes. Plus de détails sur l'objectif biologique et les données disponibles sont donnés dans la sous-section 1.1.1.

1.3 Contributions

Cette sous-section présente l'ensemble des contributions de cette thèse. Pour les problématiques soulevées en sous-section 1.2.4, on s'attardera notamment sur la question du choix de la collection de modèles, du choix de la fonction de pénalité et du choix de la calibration des constantes au sein de la pénalité. Les contributions mélangent des comparaisons de méthodes, des résultats théoriques statistiques, l'élaboration d'algorithmes pour des visées applicatives et une application sur les données réelles.

1.3.1 La sélection de variables appliquée sur un chemin de régularisation

Nous proposons d'appliquer les méthodes de sélection de modèles (les pénalités ℓ_0) à partir de la sortie des algorithmes de minimisation de critères régularisés (Lasso, Elastic-Net,..). La principale motivation est de conserver les performances prédictives obtenues sur les méthodes de sélection de modèles tout en garantissant une combinatoire raisonnable dans un contexte de grande dimension. Plus précisément, en remarquant que la fonction $\left[\lambda \mapsto \hat{\beta}_\lambda\right]$ est continue et linéaire par morceaux, nous constatons qu'un nombre fini de λ suffit pour calculer l'ensemble des estimateurs Lasso pour tout $\lambda \geq 0$. Notons Λ cette grille. Notre procédure de combinaison des deux approches se résume en trois étapes. La première consiste à résoudre le problème d'optimisation des moindres carrés régularisés (Lasso, Elastic-Net) sur Λ. La fonction $\left[\lambda \in \Lambda \mapsto \hat{\beta}_\lambda\right]$ est appelée chemin de régularisation. Pour chaque $\lambda \in \Lambda$, le support de $\hat{\beta}_\lambda$ est associé à un ensemble de variables sélectionnées et donc à un modèle

$$m = \mathrm{Vect}\Big((X_j),\ j \text{ telle que } \hat{\beta}_{\lambda,j} \neq 0\Big).$$

Puisque les estimateurs $\hat{\beta}_\lambda$ sont obtenus à partir du jeu de données disponible $\mathcal{D}$, les modèles m associés sont donc eux-aussi dépendants des données. Nous adoptons la notation $\mathcal{M}(\mathcal{D})$ pour la collection de modèles obtenus sur $\mathcal{D}$. Dans la lignée de [Efron et al., 2004, Connault, 2011], l'estimateur des moindres carrés noté $\hat{\beta}_m$ est extrait sur chaque modèle m de $\mathcal{M}(\mathcal{D})$ afin d'obtenir des estimateurs asymptotiquement sans biais. Cette étape constitue la deuxième

étape de notre procédure. Ainsi, le chemin de régularisation donne une collection de modèles $\{m \in \mathcal{M}(\mathcal{D})\}$ accompagnée d'une collection d'estimateurs $\left(\hat{\beta}_m\right)_{m\in\mathcal{M}(\mathcal{D})}$. C'est sur cette collection de modèles $\mathcal{M}(\mathcal{D})$ que sont appliquées les méthodes de sélection de modèles (les pénalités ℓ_0). Cette troisième étape repose sur la résolution du problème (1.4) où pen est une fonction positive et croissante en D_m. Le modèle sélectionné $\hat{m}$ donne l'estimateur final de β^* qui est $\hat{\beta}_{\hat{m}}$.

- Chapitre 2 :
Le chapitre 2 définit la notion de grande dimension et présente les difficultés engendrées par ce contexte. De plus, les approches décrites dans la sous-section 1.2.1 y sont détaillées. Ce chapitre 2 peut ne pas être lu par les spécialistes de la grande dimension (pour la sous-section 2.1), par les spécialistes de la sélection de modèles (pour la sous-section 2.2) et par les spécialistes des tests multiples (pour la sous-section 2.3). Contrairement à la sous-section 1.2.1 où les méthodes sont présentées dans l'ordre chronologique, elles y sont présentées sous le point de vue adopté dans cette thèse. Des détails techniques et statistiques y sont apportés. Plus précisément, le Lasso est vu comme un substitut efficace et raisonnable algorithmiquement au problème d'optimisation sous contrainte ℓ_0. Une revue non-exhaustive des propriétés théoriques du Lasso ainsi que quelques corrections du Lasso sont présentées. L'Elastic-Net, adapté aux variables corrélées, est introduit et est accompagné d'une étude comparative géométrique avec le Lasso. Grâce à la convexité des normes utilisées dans les pénalités, des algorithmes sont mis en place permettant de proposer en sortie un chemin de régularisation $\left[\lambda \in \Lambda \mapsto \hat{\beta}_\lambda\right]$. Nous présentons dans ce chapitre 2 les algorithmes *LARS* et de *descente de gradient* que nous utilisons dans cette thèse. Le chemin de régularisation est connu pour être instable, c'est pourquoi, des procédures comme Bolasso ont été proposées pour rejouer la construction du chemin de régularisation sur des réplicats du jeu de données initial, ceci afin de l'enrichir et le stabiliser. A partir de la collection de modèles $\mathcal{M}(\mathcal{D})$ obtenues et accompagnée de la collection d'estimateurs $\left(\hat{\beta}_m\right)_{m\in\mathcal{M}(\mathcal{D})}$, le problème de sélection de modèle (1.4) est considéré avec pour pen une pénalité ℓ_0. D'un côté les pénalités AIC, BIC, eBIC garantissent des propriétés asymptotiques ; d'un autre côté les pénalités proposées par L.Birgé et P.Massart et extraites d'une large étude théorique détaillée dans ce chapitre 2 garantissent des propriétés non-asymptotiques. Ces dernières sont définies par : $\forall m \in \mathcal{M},\ \mathrm{pen}(m) > \mathrm{pen}_{\min}(m)$ où $\mathrm{pen}_{\min}$ est une fonction explicite. Deux types de pénalité ont été développés dans la littérature : la pénalité *LinSelect* et les *pénalités dépendantes des données*. Dans cette thèse, nous considérons longuement ce dernier type de pénalités. Celles-ci reposent principalement sur les deux idées suivantes : les constantes inconnues apparaissant dans la pénalité sont considérées comme des hyperparamètres à calibrer directement sur le jeu de données disponible ; la variance inconnue est incluse dans les hyperparamètres. De plus, ces pénalités font intervenir une famille de poids sur la collection de modèles $\mathcal{M}(\mathcal{D})$. En adaptant cette famille à notre régression linéaire gaussienne sous nos hypothèses (grande dimension et collection de modèles aléatoire car construire sur le jeu de données), nous obtenons la pénalité de la forme suivante pour (1.4) :

$$\mathrm{pen}(m) = K\left(C_1(\sigma^2)\frac{D_m}{n} + C_2(\sigma^2)\frac{\log(\binom{p}{D_m})}{n}\right), \tag{1.6}$$

où $K > 1$, $C_1(\sigma^2)$ et $C_2(\sigma^2)$ sont les trois hyperpamètres à calibrer sur le jeu de données disponible. Ainsi, pour ce type de pénalité, la forme même de la pénalité est décidée à partir du jeu de données, en plus de la calibration des hyperparamètres.
Ce chapitre 2 s'achève sur une présentation plus détaillée et plus statistique des procédures de tests multiples. Le FDR y est introduit comme une fonction de coût permettant à la procédure de tests multiples de sélectionner des variables en s'autorisant une proportion contrôlée de sélection de variables non actives. Des alternatives au FDR ou l'introduction d'autres fonctions de coût, telle que le FNR, permettent de vérifier d'autres propriétés sur l'ensemble des variables sélectionnées.

- Chapitre 3 :
Le chapitre 3 propose une étude de comparaison de certaines procédures de sélection de variables appliquées sur des chemins de régularisation. Le cadre est la régression linéaire gaussienne avec $n = 150$ et $p = 199$. L'objectif est d'identifier les combinaisons les plus efficaces pour l'application biologique sur les données réelles. Dans un premier temps sont comparés les critères régularisés Lasso et Elastic-Net via les deux algorithmes de génération du chemin de régularisation LARS et descente de gradient. Pour cela, la métrique pROC-AUC est évaluée rendant compte de la capacité à discriminer les variables actives des variables non actives. Dans un second temps sont comparées certaines procédures de sélection de variables appliquées sur les chemins de régularisation. Parmi elles, se distinguent les procédures de sélection de modèles (les pénalités eBIC, LinSelect et deux pénalités dépendantes des données) et les procédures d'identification de variables (*ESCV*, Stability Selection, Bolasso, Tigress, et la méthode des knockoffs). Nous proposons une présentation de ces méthodes en section 3.2.
Ces comparaisons sont réalisées à partir d'une large étude de simulation comprenant des jeux de données gaussiennes où les variables sont complètement indépendantes les unes des autres et des jeux de données où les variables présentes des corrélations entre elles. La structure d'indépendance n'est pas adaptée à la pratique mais est usuellement utilisée pour développer et tester de nouveaux outils statistiques. Elle est utilisée comme référence. Pour les données corrélées, certaines sont générées via un modèle graphique gaussien sous les structures de dépendance *cluster* et *scale-free* définies en section 3.3.1, les autres sont issues d'un algorithme appelé *FRANK* et basé sur des processus dynamiques offrant des jeux de données plus réalistes et plus proches de ceux issus des réseaux de régulation de gènes. Utiliser ces données simulées corrélées est un pas intermédiaire entre l'étude des méthodes d'un point de vue théorique et leur application sur des jeux de données réelles. Pour cela, plusieurs métriques sont évaluées : le *mean squared errors (MSE)* pour les performances prédictives, la *sensibilité* (Recall) pour la capacité à sélectionner les variables actives, la *spécificité* (Specificity) pour la capacité de ne pas sélectionner les variables non actives, et le false discovery rate (FDR) pour la proportion de variables non actives parmi les variables sélectionnées.
Les conclusions sont diverses. A l'échelle des jeux de données, des dégradations presque systématiques sont observées lorsque les méthodes sont appliquées sur les variables corrélées par rapport aux variables indépendantes. Cela n'est pas surprenant puisque la théorie est souvent satisfaite lorsque des contrôles restrictifs sur la corrélation entre les variables sont vérifiés. Lorsque ceux-ci font défaut, la théorie n'est en général plus vérifiée. Les performances semblent

meilleures lorsque le support n'est pas trop petit. A l'échelle de la construction du chemin de régularisation, la combinaison Elastic-Net et LARS semble donner le chemin de régularisation tel que les variables actives soient les premières à apparaître. Au sein des méthodes de sélection de modèles, les pénalités non-asymptotiques sont à privilégier par rapport aux pénalités asymptotiques : eBIC n'est pas la meilleure méthode d'un point de vue prédictif. Cela est expliqué par la faible valeur de n. Dans ce sens, une étude plus poussée utilisant d'autres valeurs de n $(300, 600, 1200)$ met en valeur une amélioration des performances d'eBIC. Au sein des méthodes d'identification de variables, Bolasso est à privilégier par rapport à Stability Selection où les méthodes sont appliquées sur des données de plus petite tailles (de tailles $\lfloor \frac{n}{2} \rfloor$) rendant plus compliquées la sélection de variables. La pénalité LinSelect et la méthode Tigress semblent être très conservatives rendant difficile la sélection de variables.
L'application des méthodes statistiques sur les données FRANK dégrade les résultats par rapport aux données gaussiennes, et ce quelle que soit la métrique évaluée. Les méthodes de sélection de modèle perdent leurs performances de prédiction et des valeurs du FDR atteignent 1. Les ensembles de variables sélectionnées contiennent uniquement ou quasi uniquement des variables non actives. Cette étude montre que l'hypothèse gaussienne est primordiale pour ces méthodes et que l'étape de pré-processing des données est très importante si l'on veut appliquer les méthodes sur de telles données.
Il n'existe pas une meilleure méthode. En effet, le choix des méthodes à privilégier varie en fonction des métriques. D'un point de vue prédictif, ESCV et la méthode des knockoffs obtiennent les plus petites valeurs du MSE. LinSelect et eBIC sont eux aussi performants. La sensibilité et la spécificité sont deux métriques à considérer simultanément. ESCV et eBIC offrent alors les meilleurs compromis. Enfin, pour un contrôle du FDR, Bolasso, Tigress, la méthode des knockoffs et LinSelect sont performants tandis qu'ESCV est à éviter.
Ce chapitre permet de valider l'approche consistant à appliquer les méthodes de sélection de variables à partir de chemins de régularisation puisque pour chaque métrique, il existe au moins une combinaison avec des performances satisfaisantes.
Enfin, cette étude de simulation permet de révéler les mauvaises performances et les fortes instabilités obtenues pour les deux pénalités dépendantes des données, et ce pour la grande majorité des métriques et quel que soit le jeu de données utilisé. L'une des raisons principales est le choix de la famille de poids sur la collection de modèles ainsi que le choix de la calibration des hyperparamètres $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ utilisé dans cette étude. En effet, ceux-ci proviennent d'une étude expérimentale réalisée sous un autre modèle qui est la détection de rupture dans un signal. Le rapport $\frac{C_1(\sigma^2)}{C_2(\sigma^2)}$ y a été fixé à 2.5. La pénalité sous-jacente est utilisée massivement en pratique et sous d'autres modèles mais nous montrons qu'elle ne semble pas adaptée à la régression linéaire gaussienne dans un contexte de grande dimension. C'est pourquoi, les chapitres 4 et 5 s'intéressent à la calibration des hyperparamètres K (chapitre 4), $C_1(\sigma^2)$ et $C_2(\sigma^2)$ (chapitre 5) pour la pénalité (1.6) dont la forme est adaptée à notre modèle et à nos hypothèses.

1.3.2 Compromis entre le risque prédictif et le FDR en sélection de modèle

- Chapitre 4 :
Le chapitre 4 se concentre sur la calibration de l'hyperparamètre $K > 1$ en régression linéaire gaussienne dans un contexte de grande dimension. Dans ce chapitre, la collection de modèles $\mathcal{M}$ ne dépend pas des données et est complètement déterministe. Plus précisément, nous nous intéressons à la sélection de variables ordonnées : pour $q = \min(n, p)$, la collection de modèles est définie par :

$$\mathcal{M} = \Big\{ m_0 = \{0\}, m_1 = \mathrm{Vect}(X_1), m_2 = \mathrm{Vect}(X_1, X_2), \cdots, m_q = \mathrm{Vect}(X_1, X_2, \cdots, X_q) \Big\}, \tag{1.7}$$

et nous supposons que le vrai modèle $m^* = \mathrm{Vect}(X_j,\ j \text{ tel que } \beta_j^* \neq 0)$ appartient à la collection $\mathcal{M}$. En ce sens, les premières variables de la collection sont les variables actives et les suivantes ne sont que des variables non actives. Une pénalité vérifiant $\mathrm{pen} > \mathrm{pen}_{\min}$ sous ces hypothèses est : $\forall m \in \mathcal{M}$

$$\mathrm{pen}(m) = K\sigma^2 D_m, \tag{1.8}$$

avec $K > 1$, puisque $\mathrm{pen}_{\min}(m) = \sigma^2 D_m$ dans ce cadre. Elle garantit un contrôle non-asymptotique sur le risque prédictif.
Nous proposons d'étudier cette pénalité en considérant simultanément l'approche prédictive et l'approche FDR. L'objectif est de retrouver les variables actives tout en garantissant des performances en prédiction. En général, une seule des deux fonctions de coût parmi le risque prédictif et le FDR est contrôlée et cela amène à des ensembles de variables sélectionnées différents. Par exemple, dans notre contexte biologique, appliquer la sélection de modèles pour un contrôle du risque prédictif fournit un ensemble de FTs candidats et un estimateur $\hat{\beta}_{\hat{m}}$ tels que : si $(x_{n+1,1}, \cdots, x_{n+1,p})$ sont de nouveaux log-ratio de données d'expression pour les p FTs considérés, alors les valeurs de $\hat{\beta}_{\hat{m}}$ et des $(x_{n+1,1}, \cdots, x_{n+1,p})$ permettent de prédire la valeur de y_{n+1} du gène cible qui serait obtenu sous les mêmes conditions. En revanche, appliquer une méthode qui contrôle le FDR (une procédure de tests multiples par exemple) fournit un ensemble de FTs candidats tel que, à proportion d'erreur contrôlée près, la variation des valeurs des données d'expression de chaque FT entraîne celle du gène cible. Étudier ces fonctions de coût simultanément peut donner de l'information pertinente. Par exemple, si le FDR est bien contrôlé mais le risque prédictif est mauvais, alors les variables sélectionnées ont une forte probabilité d'être actives mais elles ne sont pas suffisantes pour prédire Y. Cela peut révéler la présence de variables cachées dans le modèle. Au contraire, si le risque prédictif est bien contrôlé mais le FDR mauvais, alors cela peut signifier que plusieurs groupes de variables ont des comportements similaires en terme de prédiction. Cela peut révéler la présence de fortes corrélations entre les variables.
Les procédures de sélection de modèles fournissent des ensembles de variables sélectionnées pour garantir des performances de prédiction. Or, en général, il existe de nombreuses variables dites prédictives disponibles pour construire un modèle de régression. Par conséquent, l'ensemble des variables sélectionnées contient beaucoup de variables actives mais aussi beaucoup de variables

non actives. A l'inverse, une procédure qui contrôle le FDR a tendance à être conservative : l'ensemble des variables sélectionnées contient très peu de variables inactives mais aussi peu de variables actives. Idéalement, les variables sélectionnées sont toutes les variables actives mais seulement celles-ci. C'est pourquoi, le FDR semble être la fonction de coût pertinente à ajouter dans la procédure de sélection de modèle et dans un contexte de grande dimension. Elle permet d'orienter la sélection de modèle vers la sélection de variables actives tout en conservant un contrôle sur le risque prédictif. De plus, contrairement par exemple à la combinaison FDR-FNR dont l'amélioration de l'un dégrade l'autre, obtenir un bon contrôle du risque prédictif et du FDR est envisageable. Combiner risque prédictif et FDR n'est pas nouveau mais cela n'a jamais été étudié en sélection de modèle non-asymptotique avec une pénalité ℓ_0.
Nous proposons de modifier la pénalité (1.8) de la procédure de sélection de modèle pour proposer un compromis entre le contrôle des deux fonctions de coût risque prédictif et FDR. Le seul paramètre libre est la constante K que nous proposons de faire varier. Ainsi, le critère à minimiser sur $\mathcal{M}$ devient : $\forall K > 0$, $\forall m \in \mathcal{M}$,

$$\mathrm{crit}_K(m) = ||Y - X\hat{\beta}_m||_2^2 + K\sigma^2 D_m, \tag{1.9}$$

et le modèle sélectionné $\hat{m}(K)$ est défini par :

$$\hat{m}(K) = \underset{m\in\mathcal{M}}{\arg\min}\Big\{\mathrm{crit}_K(m)\Big\}.$$

D'un point de vue théorique, toute constante $K > 1$ permet d'obtenir un contrôle non-asymptotique sur le risque prédictif. En pratique, K est souvent fixée à 2 car 2 est la constante optimale d'un point de vue prédictif et asymptotique. En revanche, il n'existe pas de constante optimale universelle pour une considération non-asymptotique. Sur la Figure1.5 sont tracées les courbes des estimations empiriques $\mathrm{FDR}\big(\hat{m}(K)\big)$ et $\mathrm{PR}\big(\hat{m}(K)\big)$ pour tout $K > 0$. Des constantes proches de 2 donnent des performances prédictives similaires à celles obtenues pour $K = 2$. En ce qui concerne le FDR, plus K augmente, plus le FDR diminue. Ainsi, un $K > 2$ bien choisi permet de conserver les performances de prédiction garanties par la sélection de modèles et d'ajouter un contrôle plus strict sur le FDR. Dans cette direction, nous proposons d'étudier la fonction $\Big[K > 0 \mapsto \mathrm{FDR}(\hat{m}(K))\Big]$ en sélection de modèles.
Bien que la procédure de sélection de modèles est construite pour des considérations de prédiction, nous encadrons la fonction $\Big[K > 0 \mapsto \mathrm{FDR}(\hat{m}(K))\Big]$ pour tout (n, p) fixés et en supposant que σ^2 est connu :

$$b(K, \beta^*, \sigma^2) \leq \mathrm{FDR}(\hat{m}(K)) \leq B(K, \beta^*, \sigma^2)$$

où $\Big[K > 0 \mapsto b(K, \beta^*, \sigma^2)\Big]$ et $\Big[K > 0 \mapsto B(K, \beta^*, \sigma^2)\Big]$ sont sous formes explicites et facilement implémentables. En effet, elles ne font intervenir que certaines évaluations des fonctions de répartition de la loi gaussienne et de certaines loi du χ-2. De plus, elles ne dépendent pas des données. Une illustration des bornes est proposée Figure1.6. Pour obtenir ces bornes, l'idée clé est de ré-exprimer la quantité $\mathrm{FDR}(\hat{m}(K))$ au sein de la sélection de modèle, c'est-à-dire faisant intervenir la définition de $\hat{m}(K)$ comme minimiseur du critère (1.9). Nous menons une étude asymptotique en K et nous prouvons la convergence vers 0 avec une vitesse exponentielle

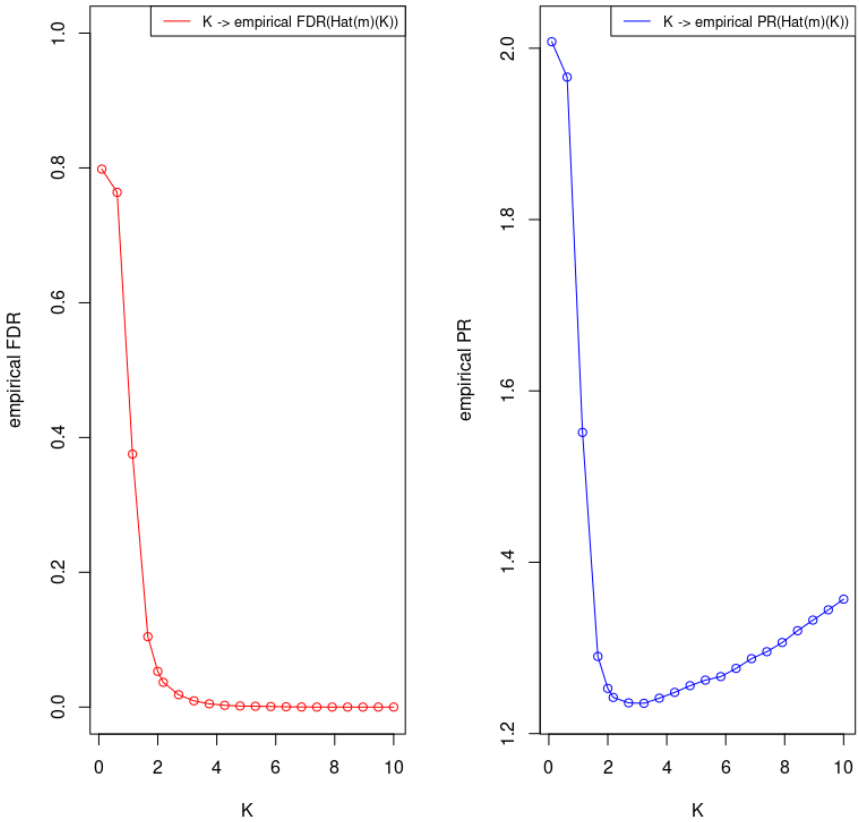


Figure 1.5: Courbes des estimations empiriques de $\mathrm{FDR}\big(\hat{m}(K)\big)$ et $\mathrm{PR}\big(\hat{m}(K)\big)$ pour $K > 0$.

des bornes $\Big[K > 0 \mapsto b(K, \beta^*, \sigma^2)\Big]$ et $\Big[K > 0 \mapsto B(K, \beta^*, \sigma^2)\Big]$. Cela montre que les bornes sont proches du terme $\mathrm{FDR}(\hat{m}(K))$ très rapidement lorsque K augmente.
En revanche, les termes $b(K, \beta^*, \sigma^2)$ et $B(K, \beta^*, \sigma^2)$ dépendent des paramètres du modèle inconnus β^* et σ^2. Pour une considération pratique où β^* et σ^2 sont inconnus, nous montrons, à partir d'une large étude de simulation, que σ^2 peut raisonnablement être estimé par $\hat{\sigma}^2$ via la *méthode de l'heuristique de pente* et que β^* peut être remplacé par $\hat{\beta}_{\hat{m}(4)}$ pour la borne supérieure. En ce qui concerne le risque prédictif, nous proposons l'estimé suivant pour chaque $K > 0$:

$$\widehat{\mathrm{PR}}(\hat{m}(K)) = \frac{1}{n}\sum_{i=1}^{n}\Big(\big(X\hat{\beta}_{\hat{m}(2)}\big)_i - \big(X\hat{\beta}_{\hat{m}(K)}\big)_i\Big)^2. \tag{1.10}$$

Elle permet d'évaluer les performances de prédiction de $\hat{m}(K)$ par rapport à la référence obtenue avec $\hat{m}(2)$. Cette quantité (1.10) est complètement disponible sur un jeu de données et ne nécessite pas de séparer les données en deux (classiquement : un jeu pour l'entraînement et un jeu pour la validation).
Les résultats théoriques nous permettent, à partir de la quantité (1.10) et de la borne supérieure

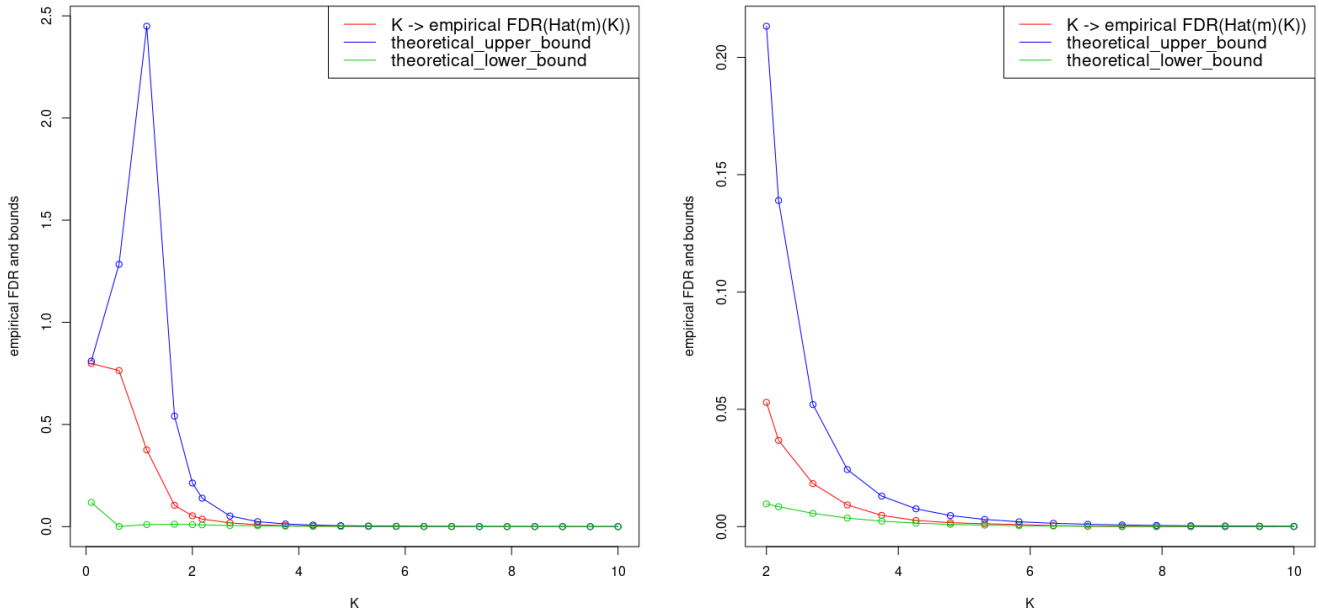


Figure 1.6: Courbes de l'estimation empirique du FDR et des termes $b(K, \beta^*, \sigma^2)$ et $B(K, \beta^*, \sigma^2)$ lorsque la matrice X est orthogonale. Droite : Zoom pour une meilleure lisibilité : les courbes ne sont tracées que pour $K \geq 2$.

estimée $\tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$, de proposer un algorithme complètement dépendant des données pour la calibration de l'hyperparamètre K dans la pénalité (1.8) :

1. Choisir α pour contrôler le FDR.
2. Calculer $I_1 = \left\{K \geq 2,\ \tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2) \ \in]0, \alpha[\right\}$.
3. Calculer $I_2 = \left\{K \geq 2,\ \widehat{\mathrm{PR}}(\hat{m}(K)) \approx \widehat{\mathrm{PR}}(\hat{m}(2))\right\}$.
4. Si $I_1 \cap I_2 \neq \emptyset$, choisir $\min\left\{K,\ K \in I_1 \cap I_2\right\}$. Sinon, choisir $\min\left\{K, K \ \in I_1\right\}$.

Cet algorithme est validé via une large étude expérimentale : le K obtenu permet d'assurer une faible valeur sur les estimations empiriques du risque prédictif et du FDR.
La collection de modèles utilisée (1.7) et l'hypothèse $m^* \in \mathcal{M}$ ne permettent pas d'appliquer notre algorithme sur les données réelles sans travail supplémentaire. Étendre les résultats à une collection de modèles $\mathcal{M}$ quelconque voire à une collection de modèles $\mathcal{M}(\mathcal{D})$ dépendante des données est une étape intermédiaire nécessaire. Elle n'est pas traitée dans cette thèse.

1.3.3 L'heuristique de pente en dimension 2

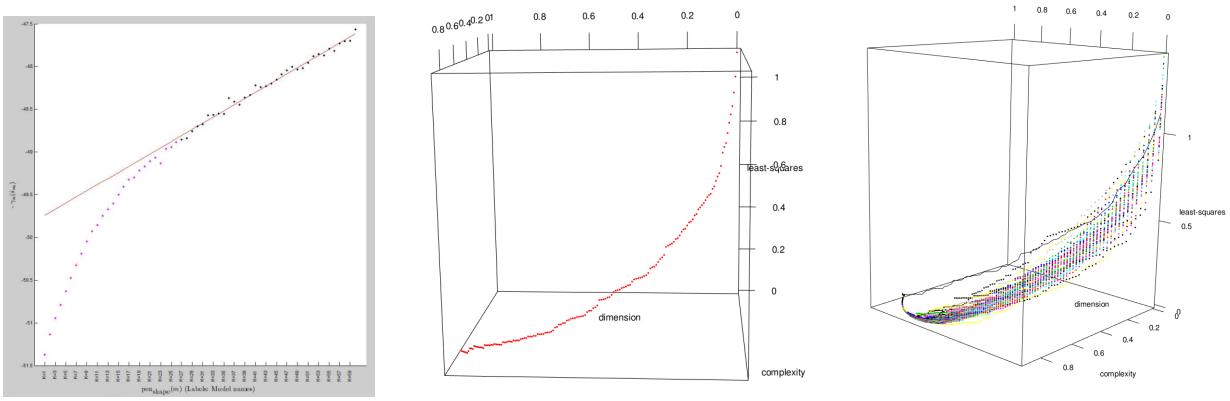
- Chapitre 5 :
Nous avons vu au chapitre 3 les mauvaises performances des pénalités dépendantes des données. Ainsi, le chapitre 5 propose un nouveau critère de sélection de modèle non-asymptotique et construit sur les données, ceci par une généralisation de la méthode de l'heuristique de pente en régression linéaire gaussienne dans un contexte de grande dimension. Dans ce chapitre, la variance σ^2 est supposée inconnue et la collection de modèles $\mathcal{M}(\mathcal{D})$ est aléatoire et dépendante des données.
Le principe de l'heuristique de pente a été introduit pour pallier le problème de constantes inconnues apparaissant dans les fonctions de pénalité dépendantes des données (voir chapitres 2 et 3 pour une introduction à ces pénalités). Les constantes sont vues comme des hyperparamètres qui sont calibrés à partir du jeu de données disponible. Il s'agit d'une heuristique mais qui est fondée sur les résultats théoriques de L.Birgé et P.Massart que nous présentons dans le chapitre 5. En effet, les pénalités pen dans (1.4) vérifiant pen $>$ pen_{min} se comportent théoriquement comme une fonction affine en les moindres carrés lorsque les dimensions de modèles D_m sont assez grandes. Les constantes multiplicatives du comportement affine permettent de calibrer les hyperparamètres de pen. A ce jour, l'heuristique de pente n'existe qu'en dimension 1, c'est-à-dire pour pour calibrer un seul hyperparamètre. De plus, elle considère une collection de modèles $\mathcal{M}$ déterministe. La méthode est implémentée dans la fonction R *DDSE* du package R *Capushe* et se compose de trois étapes :

1. Simplifier la collection de modèles $\mathcal{M}$ disponible.
2. Estimer des constantes sur plusieurs comportements affines entre les valeurs des moindres carrés et les valeurs de la fonction de pénalité.
3. Rechercher le meilleur comportement linéaire et sélectionner la constante associée.

Lorsque plusieurs hyperparamètres existent dans la pénalité, la méthode de l'heuristique de pente est utilisée soit en appliquant l'heuristique de pente de dimension 1 successivement sur chacune des constantes et en fixant toutes les autres ; soit en utilisant le rapport 2.5 (voir le chapitre 3 pour plus de détails). Cette dernière approche est incluse dans l'étude de comparaison réalisée dans le chapitre 3 qui révèle de mauvaises performances. Une raison peut être que le rapport $\frac{C_1(\sigma^2)}{C_2(\sigma^2)}$ n'est pas égal à 2.5 et n'est pas nécessairement fixe en régression linéaire gaussienne en grande dimension. La pénalité sous-jacente n'est pas adaptée, ce qui peut engendrer l'inexistence d'un comportement affine entre les moindres carrés et la pénalité. L'heuristique de pente en dimension 1 y est donc inefficace et la calibration de l'hyperparamètre est mauvaise, ainsi que la sélection de modèles.
Nous proposons d'améliorer l'algorithme de l'heuristique de pente et dans notre cadre, la pénalité est (1.6) où K est fixé à 2 pour ce chapitre. Deux hyperparamètres $\Big(C_1(\sigma^2), C_2(\sigma^2)\Big)$ sont donc à calibrer. Nous proposons de répondre à deux objectifs.
Le premier est la calibration de $\Big(C_1(\sigma^2), C_2(\sigma^2)\Big)$. Pour cela, notre algorithme s'inspire de la

fonction R *DDSE* du package R *Capushe* mais l'étape 2 est modifiée. A la place des régressions linéaires robustes utilisées, nous ré-écrivons le problème d'estimation des constantes comme un problème de minimisation d'un critère matriciel sous contraintes de positivité des deux constantes. Pour cela, la fonction R *solve.QP* du package R *quadprog* remplace la fonction R *rlm*. Ainsi, les constantes obtenues sont positives, ce qui est un gain par rapport à l'heuristique de pente en dimension 1, mais cela a un prix : la perte des propriétés de robustesse que vérifie l'heuristique de pente en dimension 1. Une illustration du comportement des moindres carrés en fonction de la pénalité est proposée Figure1.7 à gauche en dimension 1 et au milieu en dimension 2.
Le deuxième objectif est d'adapter l'heuristique de pente à une collection de modèles aléatoire $\mathcal{M}(\mathcal{D})$. Pour cela, l'étape 1 de l'algorithme est modifiée par l'ajout d'une procédure *Bootstrap* en rejouant la création de la collection de modèles sur des ré-échantillons du jeu de données initial. La collection de modèles utilisée lors des étapes 2 et 3 est alors enrichie. Une illustration de différents chemins de régularisation obtenus par l'utilisation de ré-échantillons est proposée Figure1.7 à droite.

Les deux algorithmes proposés sont testés sur les mêmes jeux de données *cluster* et *scale-free*



1D 2D 2D avec ré-échantillonage

Figure 1.7: Exemples des valeurs de $-\gamma_n\left(\hat{\beta}_m\right)_{m\in\mathcal{M}(\mathcal{D})}$ en fonctions des valeurs de la pénalité sur la ou les (courbe de droite) collection(s) de modèles [3].

que le chapitre 3 et les modèles sélectionnés sous-jacents sont mis en comparaison avec ceux issus de l'heuristique de pente en dimension 1 et ceux issus de la pénalité LinSelect. La comparaison avec la pénalité LinSelect est importante puisque cette dernière garantit une inégalité oracle non-asymptotique sur le risque prédictif sous les hypothèses de notre cadre statistique : grande dimension, variance inconnue, collection de modèles aléatoire $\mathcal{M}(\mathcal{D})$. Cependant cette fonction de pénalité est non flexible car totalement déterministe. Au contraire, les pénalités dépendantes des données sont flexibles car elles s'adaptent aux données lors de la calibration des constantes mais elles ont été construites pour des collections de modèles déterministes.

[3]La figure est extraire de [Baudry et al., 2012]

Ainsi, nos algorithmes proposent une procédure dépendante des données à la fois à l'échelle de la collection de modèle (comme pour LinSelect) et à la fois à l'échelle de la sélection du modèle (comme pour les pénalités dépendantes des données).
Les algorithmes proposés diffèrent de l'heuristique de pente en dimension 1 : le rapport $\frac{C_1(\sigma^2)}{C_2(\sigma^2)}$ n'est pas fixe et non égal à 2.5. De plus, lorsque le ré-échantillonge est effectué, le risque prédictif des modèles sélectionnées par nos algorithmes est plus petit que ceux issus de l'heuristique de pente en dimension 1 et de l'utilisation de la pénalité LinSelect.
Ces deux algorithmes peuvent être appliqués sur les données réelles mais, contrairement au chapitre 3, seul le risque prédictif est étudié sur ces nouvelles méthodes.

1.3.4 Proposer des FTs candidats d'un gène cible à partir de données transcriptomiques

- Chapitre 6 :
Les travaux de cette thèse sont en particulier motivés par l'étude de données transcriptomes. Le chapitre 6 a pour objectif de répondre à la question biologique suivante : peut-on retrouver les *facteurs de transcription* impliqués dans le mécanisme de régulation d'un gène cible à partir de données transcriptomiques? Pour cela, nous nous intéressons à quatre gènes cibles d'*Arabidopsis thaliana*. Ces derniers, qui sont également des *facteurs de transcription*, sont massivement étudiés et sont connus pour interagir physiquement ensemble. De plus, une liste de *facteurs de transcription* basée sur de la connaissance biologique est disponible pour chacun des gènes cibles. Cette liste tient lieu de référence. Manipuler un jeu de données réelles nécessite une étape incontournable de pré-processing. Ainsi, nous proposons dans un premier temps un traitement des données manquantes, puis nous montrons qu'un comportement linéaire entre les gènes peut être visible. Enfin, l'hypothèse gaussienne n'est pas tout à fait vérifiée sur les données réelles : la distribution des données par gène est à décroissance exponentielle plus rapide que la gaussienne. C'est pourquoi, nous préférons conserver les données brutes plutôt que d'y appliquer une quelconque transformation puisque le but n'étant pas de minimiser l'erreur d'approximation du modèle sur les données mais d'y appliquer et de comparer différentes méthodes statistiques. Plus précisément, sont appliquées les pénalités eBIC, LinSelect, une pénalité dépendante des données ainsi que les méthodes Bolasso, Tigress et les knockoffs, toutes considérées dans l'étude de comparaison du chapitre 3. Les algorithmes proposés dans le chapitre 5 y sont également testés. Les différentes procédures statistiques se comportement de manière similaire en sélectionnant la plupart du temps les mêmes *facteurs de transcription*. Cependant, la plupart des *facteurs de transcription* présents dans la liste de référence ne sont pas sélectionnés. L'analyse proposée ne permet pas de répondre précisément à la question biologique mais soulève plusieurs conclusions possibles : la connaissance biologique est parcellaire et les *facteurs de transcription* connus ne sont pas les acteurs majeurs ; la modélisation statistique n'est pas adaptée à la question biologique (à travers l'étape de pré-processing ou les méthodes statistiques testées) ; les données transcriptomiques ne permettent pas de répondre à la question biologique.

Guide de lecture

Cette thèse est divisée en 5 parties distinctes. Le chapitre 2 peut ne pas être lu par les spécialistes de la grande dimension (pour la sous-section 2.1), par les spécialistes de la sélection de modèles (pour la sous-section 2.2) et par les spécialistes des tests multiples (pour la sous-section 2.3). Le chapitre 4 peut être lu indépendemment des autres. Le chapitre 3 fait un état de l'art des différentes procédures possibles en sélection de variables. Les résultats de leur comparaison par une large étude de simulation sont un moteur à la création du nouveau critère de sélection de modèle non-asymptotique et construit sur les données que nous proposons dans le chapitre 5. Enfin, le chapitre 6 applique certaines méthodes statistiques décrites dans le chapitre 3 ainsi que le nouveau critère du chapitre 5 sur les données réelles. Il répond de manière parcellaire à la problématique biologique de cette thèse.

Un ensemble de perspectives est proposé en conclusion de cette thèse. Elles sont plus générales que celles proposées à la fin de chacun des autres chapitres.

Les preuves et les calculs techniques des chapitres 4 et 5 sont reportés dès que possible dans une sous-section à part ou en annexe.

Les chapitres 3 et 4 sont écrits sous forme d'articles qui seront soumis très prochainement. Les autres chapitres sont écrits sous forme de chapitres de thèse.

Le chapitre 3 a été réalisé en collaboration avec Mélina Gallopin de l'Institut de Biologie Intégrative de la Cellule (I2BC). Le chapitre 6 est le fruit d'un encadrement de projet de deux étudiants de Master 2 "Mathématiques pour les Sciences du Vivant" de l'Université Paris-Saclay : Marion Naveau et Armand Favrot.

Nous avons disposé des ressources informatiques de la plateforme Migale de l'INRAE (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi : 10.15454/1.5572390655343293E12) pour faire tourner les codes utiles pour les chapitres 3, 5 et 6.

Chapter 2

Des outils statistiques pour la sélection de variables

Résumé

Ce chapitre 2 détaille dans un premier temps l'impact de la *grande dimension* sur l'analyse de données (sous-section 2.1). Puis, il détaille deux approches différentes à la procédure de sélection de variables utilisées en régression linéaire gaussienne de grande dimension. Ces approches contrôlent deux fonctions de coût différentes : le *risque prédictif (RP)* (sous-section 2.2) et le *False Discovery Rate (FDR)* (sous-section 2.3). La première est l'approche prédictive via une procédure de pénalisation de critères pour un contrôle du risque prédictif en moyenne. La seconde est l'approche sélection via une procédure de tests multiples pour un contrôle de quantiles et du FDR. La première répond à un problème d'estimation alors que la seconde répond à un problème de test.

Le chapitre 2 peut ne pas être lue par les spécialistes de la grande dimension (pour la sous-section 2.1), par les spécialistes de la sélection de modèles (pour la sous-section 2.2) et par les spécialistes des tests multiples (pour la sous-section 2.3).

Contents

2.1 La grande dimension

La grande dimension impacte la procédure statistique de diverses manières [Giraud, 2014]. Premièrement, au plus la dimension d'un espace grossit, au plus les données disponibles sont isolées les unes par rapport aux autres : les données les plus proches sont en fait très éloignées dans l'espace. La Figure 2.1 illustre ce phénomène : 12 points sont aléatoirement placés respectivement sur le segment $[0,1]$, le carré $[0,1]^2$ et le cube $[0,1]^3$. Au plus la taille de l'espace grandit, au plus il est difficile de définir une proximité entre les points. En fait, pour avoir la distance euclidienne plus petite que 0.01 entre tous les points, il faut 100, 10000 et 10^{20} points équi-répartis dans respectivement $[0,1]$, $[0,1]^2$ et $[0,1]^{10}$: la croissance est exponentielle en la taille de l'espace. L'impact de la grande dimension va au delà de ce phénomène et est non intuitif puisque dans un espace de grande dimension, tous les points d'un nuage de points quelconque sont à une distance similaire les uns des autres et se concentrent sur une sous-variété de faible dimension. Des structures ou des similarités sont donc difficiles à détecter au sein de ce nuage. Ce phénomène s'appelle le *fléau de la grande dimension*. Par exemple, des points placés uniformément au hasard dans $[-1,1]^d$ pour d relativement grand se retrouvent concentrés près de la sphère de rayon $\sqrt{\frac{d}{3}}$. Cela est une conséquence de la propriété : $\frac{\mathrm{Vol}\left(B(0,1)^d\right)}{\mathrm{Vol}\left([-1,1]^d\right)} \underset{d\longrightarrow+\infty}{\longrightarrow} 0$ où $B(0,1)^d$ désigne la boule ouverte unitaire de $\mathbb{R}^d$. Deuxièmement, la grande dimension impacte la fluctuation globale d'une estimation car celle-ci est la somme des fluctuations coordonnée par coordonnée. Par exemple, dans le cas de notre modèle de régression linéaire gaussienne (1.1), l'erreur quadratique entre $\hat{\beta}$ et β^* pour un estimateur $\hat{\beta}$ quelconque peut être grande puisque le nombre de composantes dans β^* est grand. Cela est causé par le fait que le maximum de la densité Gaussienne décroît exponentiellement vite avec la dimension, ainsi, la queue de distribution est de plus en plus chargée. Troisièmement, les événements rares s'accumulent avec la

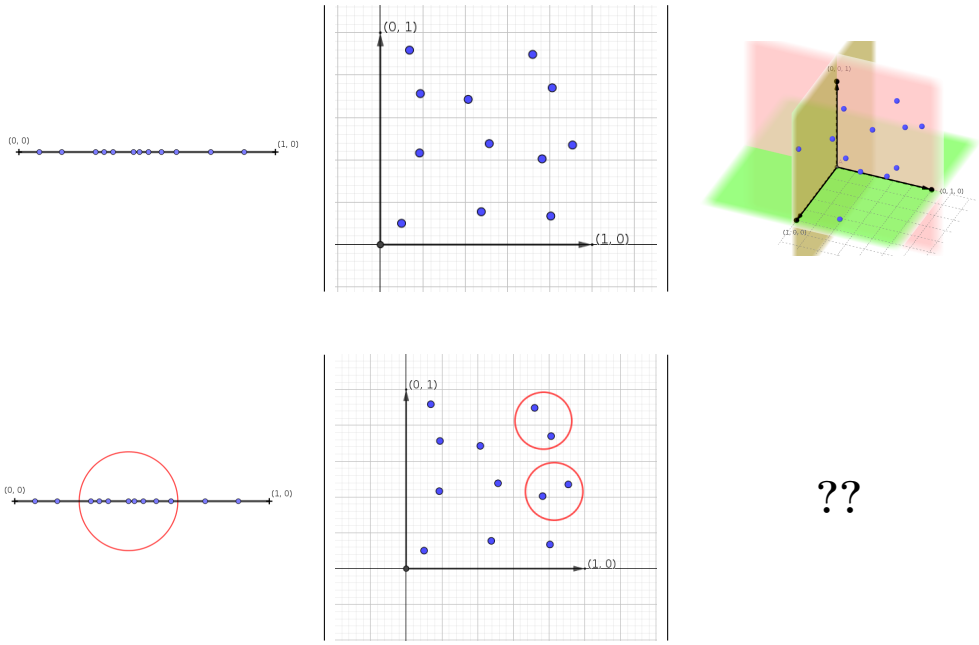


Figure 2.1: Illustration du fléau de la grande dimension. 12 points sont aléatoirement placés respectivement sur le segment $[0,1]$ (à gauche), le carré $[0,1]^2$ (au milieu) et le cube $[0,1]^3$ (à droite).

croissance de la taille de l'espace, ce qui peut fausser l'estimation. Quatrièmement, les calculs numériques explosent le temps computationnel possible et la mémoire disponible pour des données de grande dimension. Dans le cas de notre modèle de régression linéaire gaussienne (1.1), retrouver les variables impliquées pour expliquer Y revient à retrouver le meilleur sous-ensemble de $\{1, \cdots, p\}$ correspondant au support de β^*. Le nombre de tel sous-ensemble est de 2^p. La valeur de p n'a pas besoin d'être très grande pour rendre l'exploration exhaustive de chacun d'eux computationnellement impossible. Par exemple, pour $p = 15$, alors il y a $2^p = 32768$ sous-ensemble à explorer.

Notons que le nombre de variables p pris en compte dans le modèle de régression linéaire gaussienne (1.1) est la dimension de l'espace dans lequel est estimé le paramètre inconnu β^* (qui est $\mathbb{R}^p$). En plus des quatre problèmes cités dans le paragraphe précédent, cette grande valeur de p influe sur le nombre d'observations n nécessaires pour la résolution du problème statistique : si p est grand, alors n doit l'être aussi : ce phénomène est appelé le *Big Data*. La Figure 2.2 illustre ce problème pour $p = 1$ et $p = 2$ dans le cadre de la régression linéaire déterministe. Si $p = 1$, alors l'unique paramètre à trouver est $\beta_1 \in \mathbb{R}$ tel que $Y = \beta_1 X_1$ (recherche d'une droite passant par l'origine). Une unique observation (y_1, x_1) est suffisante pour donner l'existence et l'unicité du β_1. Si $p = 2$, les paramètres à trouver sont (β_1, β_2) tels que $Y = \beta_1 X_1 + \beta_2 X_2$ (recherche d'un plan passant par l'origine). Dans ce cas, deux observations $\Big((y_1, x_1), (y_2, x_2)\Big)$ sont nécessaires pour donner l'existence et l'unicité du couple (β_1, β_2), une seule observation donne une infinité de solutions (trois exemples de plan sont représentés en bas de la Figure 2.2). En généralisant pour un p quelconque, $n = p$ observations

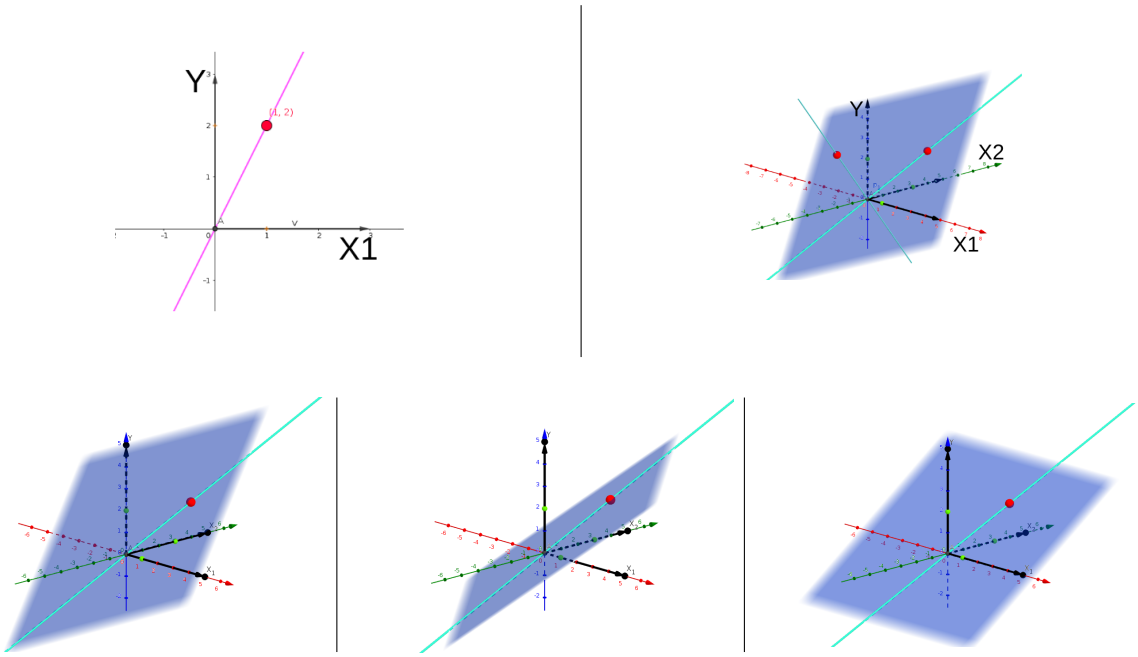


Figure 2.2: Illustration du problème de l'estimation des paramètres dans une régression linéaire déterministe pour $p = 1$ et $p = 2$. En haut, $n = 1$ observation est nécessaire (respectivement $n = 2$ observations sont nécessaires) pour avoir une unique droite (respectivement plan) passant par l'origine et le point (respectivement les deux points). En bas, trois exemples de plan sont donnés passant par l'origine et un point.

sont nécessaires pour donner l'existence et l'unicité de $(\beta_1, \cdots, \beta_p)$ tel que $Y = \beta_1 X_1 + \cdots \beta_p X_p$. Lorsque de l'aléa est ajouté, la tendance linéaire est à trouver à travers un nuage de points. Plus d'observations sont nécessaires pour assurer de bonnes propriétés d'estimation. Ainsi, même si n est très grand devant p, sa valeur peut ne pas être suffisante pour obtenir des garanties (cela dépend de la valeur de p).

Pour conclure, le cadre de la grande dimension est large et varié. Il intervient lorsque p est grand mais aussi lorsque la valeur de p est proche ou dépasse la valeur de n.

2.2 Minimisation du risque prédictif via des méthodes pénalisées

Pénaliser un critère d'ajustement linéaire est une méthode pour sélectionner des variables. Elle consiste à imposer la parcimonie par l'introduction d'une fonction de pénalité dans le critère d'ajustement linéaire à optimiser. Cette méthode est divisée en deux étapes : la première est la génération d'une collection de modèles contenant quelques estimateurs pertinents (sous-section 2.2.1) ; la seconde est la sélection du meilleur estimateur parmi la collection disponible pour un contrôle du risque prédictif. Ces deux étapes forment une procédure de sélection de

modèle (sous-section 2.2.2).

2.2.1 La collection de modèle

Notons $||.||_{2,n}$ la norme euclidienne normalisée usuelle sur $\mathbb{R}^n$. Pour $q \in \{0,1\}$, $|.|_q$ désigne la norme q usuelle sur $\mathbb{R}^p$ et pour $q \geq 2$ entier, $||.||_q$ désigne la norme q usuelle sur $\mathbb{R}^p$. Ces normes sur $\mathbb{R}^p$ sont par la suite appelées les normes ℓ_q. Soit

$$m^* = \text{Vect}(X_j,\ j \text{ tel que } \beta_j^* \neq 0)$$

le sous-espace vectoriel de $\mathbb{R}^p$ engendré par les variables actives. Si m^* est connu, alors l'estimateur des moindres carrés dans ce sous-espace vectoriel m^* défini par :

$$\hat{\beta} = \underset{\{\beta \in \mathbb{R}^p, X\beta \in m^*\}}{\arg\min} ||Y - X\beta||_{2,n}^2$$

donne le meilleur vecteur $X\hat{\beta}$ pour prédire Y.
Cependant, m^* est inconnu en pratique. De plus, comme aucun ordre naturel sur les variables n'est donné en général, tous les sous-ensembles de variables possibles doivent être explorés pour retrouver m^*. Cela représente 2^p ensembles en toute généralité. Dans un contexte de grande dimension, une telle exploration est impossible. De plus, il est nécessaire de prendre en compte l'hypothèse de parcimonie dans la procédure pour empêcher la sélection d'un sous-ensemble trop gros. Ainsi, une alternative est de contraindre l'estimateur des moindres carrés à avoir un support petit. Dans cette direction, il est naturel de proposer la minimisation sous contrainte suivante : pour $t > 0$,

$$\hat{\beta}_t = \underset{\{\beta \in \mathbb{R}^p, |\beta|_0 \leq t\}}{\arg\min} ||Y - X\beta||_{2,n}^2$$

où $|\beta|_0 := \#\{j, \beta_j \neq 0\}$. Par cette contrainte, $\hat{\beta}_t$ a un support de taille au plus t. Tout l'enjeu est de trouver la bonne valeur de $t > 0$: au plus t est petit, au plus la parcimonie est forte mais au moins l'ajustement du modèle sur les données est bon ; au plus t est grand, au plus l'ajustement du modèle sur les données est bon mais au moins la parcimonie est forte. Une bonne valeur de t réalise un équilibre entre la qualité de l'ajustement du modèle sur les données et le respect de l'hypothèse de parcimonie. Ré-écrit sous sa forme Lagrangienne [Ismaili and Gaillard, 2009], cette minimisation sous contrainte devient la minimisation du critère suivant : pour $\lambda > 0$,

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \Big\{ ||Y - X\beta||_{2,n}^2 + \lambda |\beta|_0 \Big\} \tag{2.1}$$

Trouver le bon t est équivalent à trouver le bon paramètre $\lambda \geq 0$ dans (2.1) : pour $\lambda = 0$, l'estimateur $\hat{\beta}_\lambda$ correspond à l'estimateur des moindres carrées, son support est plein et toutes les variables sont sélectionnées ; plus λ augmente, plus le nombre de variables sélectionnées diminue et moins l'ajustement du modèle sur les données est bon ; pour λ assez grand, la parcimonie est totale et aucune variable n'est sélectionnée. Par la contrainte sur le nombre de coordonnées non nulles sur β, la minimisation de l'équation (2.1) conduit à une sélection

de variables : seules les variables indexées par j pour les j tels que $\hat{\beta}_{\lambda,j} \neq 0$ sont considérées comme actives.

L'ajout de la contrainte ℓ_0 (2.1) fait perdre à la minimisation des moindres carrés la convexité et la dérivabilité sur $\mathbb{R}^n$ du critère à minimiser. Algorithmiquement, l'estimateur associé est incalculable dès que p est de l'ordre de quelques dizaines. Une idée est de trouver une autre pénalisation, proche de la norme ℓ_0, afin de rendre le problème d'optimisation sous contrainte accessible algorithmiquement en grande dimension. En remarquant que :

$$|\beta|_0 = \sum_{j=1}^{p} \mathbb{1}_{\beta_j \neq 0} = \lim_{q \to 0} \sum_{j=1}^{p} |\beta_j|^q = \lim_{q \to 0} ||\beta||_q^q,$$

la norme ℓ_0 peut être remplacée par une norme ℓ_q avec $q > 0$ proche de 0. Si $q < 1$, ℓ_q n'est pas une norme ; si $q \geq 1$, ℓ_q est une norme et en plus elle est convexe.

- Le Lasso :

Dans son article, [Tibshirani, 1996] propose la norme ℓ_1 comme substitut efficace à la norme ℓ_0, et on obtient l'estimateur Lasso (Least Absolute Shrinkage and Selection Operator) en minimisant le critère suivant pour $\lambda > 0$,

$$\hat{\beta}_{\lambda,\text{Lasso}} = \arg\min_{\beta \in \mathbb{R}^p} \Big\{ ||Y - X\beta||_{2,n}^2 + \lambda |\beta|_1 \Big\}$$

Pour un $\lambda > 0$ fixé, on peut montrer que le Lasso estime à zéro les coefficients de $\hat{\beta}_{\lambda,\text{Lasso}}$ en dessous d'un certain seuil qui dépend de λ. Les variables correspondantes sont donc retirées du support, ainsi, le Lasso sélectionne automatiquement des variables. La dépendance en le paramètre λ est donc forte et son choix est un réel challenge mais remarquons que la fonction $\big[\lambda \mapsto \hat{\beta}_{\lambda,\text{Lasso}}\big]$ est continue et linéaire par morceaux. Ainsi, pour calculer l'ensemble des estimateurs Lasso pour tout $\lambda \geq 0$, seuls un nombre fini de λ suffit. Cette grille finie de valeurs de paramètres de régularisation λ donne l'ensemble des sous-ensembles de variables qu'explore le Lasso. Cet ensemble forme le *chemin de régularisation* du Lasso. Ainsi, Lasso évince donc la plupart des 2^p ensembles possibles pour ne garder que les plus pertinents.

Pour tout $\lambda > 0$, il existe toujours une solution au problème Lasso, celle-ci n'étant pas forcément unique mais la prédiction $X\hat{\beta}_{\lambda,\text{Lasso}}$ l'est. Les résultats théoriques sur le Lasso sont nombreux et variés (performance de prédiction, qualité de l'estimation, sélection de variables, en tant qu'algorithme de régularisation). Tout d'abord, une inégalité oracle sur $||X\beta^* - X\hat{\beta}_{\lambda,\text{Lasso}}||_{2,n}^2$ pour $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}$ avec $A > 2\sqrt{2}$ est vérifiée avec grande probabilité sous une condition de valeur propre restreinte [Bickel et al., 2009]. Deuxièmement, est traitée dans [Bickel et al., 2009] la question légitime de l'estimation du paramètre puisque le Lasso met à 0 tous les petits coefficients, ce qui a un impact a priori sur l'estimation des coefficients restants. Sous la même condition de valeur propre restreinte que précédemment, une inégalité sur $||\beta^* - \hat{\beta}_{\lambda,\text{Lasso}}||_{2,n}^2$ est satisfaite avec grande probabilité. Malheureusement, cette condition sur les valeurs propres n'est, en pratique, pas satisfaite en grande dimension. Troisièmement, la capacité du Lasso à

sélectionner les bonnes variables a été étudiée. Elle est en général vérifiée si une condition, appelée dans la littérature *condition d'irreprésentabilité (CI)* est vérifiée. Cette condition contrôle la corrélation entre chaque variable actives avec chacune des variables non actives. Sous une condition similaire à CI, alors la *consistance en sélection asymptotique* (probabilité de sélectionner les bonnes variables tend vers 1 lorsque n tend vers l'infini) de l'estimateur Lasso est vérifiée [Zou, 2006,Meinshausen et al., 2006]. Cependant, dès que CI est violée, alors le Lasso est inconsistant [Zou, 2006, Zhao and Yu, 2006]. La raison est l'estimation brutale à 0 pour toutes les composantes petites de l'estimateur, ainsi, s'il existe de fortes corrélations entre les variables actives et les autres (CI non respectée), alors le Lasso a tendance à sélectionner aléatoirement des variables non impliquées dans la régression mais fortement corrélées à celles impliquées dans la régression. Ainsi, si CI est violée, le Lasso a tendance à sélectionner trop de variables. Par ailleurs, la condition CI est suffisante pour la *consistance en signe non-asymptotique* (la probabilité que les signes des coefficients soient les bons est grande) [Zhao and Yu, 2006]. La quasi nécessité de CI a été démontrée si $p \leq n$ pour obtenir la *consistance en signe asymptotique*. Enfin, des inégalités oracles sur $||X\hat{\beta}_\lambda - X\beta^*||_{2,n}^2 + \lambda|\hat{\beta}_\lambda|_1$ ont été obtenus dans [Huang, 2008, Rigollet and Tsybakov, 2011, Bartlett et al., 2012, Massart and Meynet, 2010] permettant d'évaluer la qualité du Lasso vu comme un algorithme de régularisation. Notons que cette condition CI est difficile à vérifier en pratique et qu'elle peut facilement ne pas être remplie en présence de corrélations entre les variables [Meinshausen et al., 2009].

- Versions modifiées du Lasso :
Pour pallier le problème de sélectionner trop de variables, plusieurs corrections du Lasso ont été proposées : le *Lasso adaptatif* [Zou, 2006] consistant à mettre des poids adaptatifs sur les coordonnées de β lors de la construction du chemin de régularisation, ou encore le *Lasso seuillé* [Van de Geer et al., 2011] qui ne conserve que les variables correspondantes à une coordonnée supérieure à un certain seuil.

- Vers d'autres pénalités : pénalités Ridge et Elastic-Net
Une alternative au seuillage brutal à 0 des petits coefficients de l'estimateur Lasso est d'utiliser d'autres valeurs de $q \geq 1$ pour la norme ℓ_q. En ce sens, l'estimateur *Ridge* a été introduit par [Hoerl and Kennard, 1988] et utilise la pénalité ℓ_2 pour construire le chemin de régularisation. Cette pénalité est convexe et dérivable sur $\mathbb{R}^n$, ainsi, elle permet une certaine régularité dans l'estimation des coefficients : les petites valeurs ne sont plus estimées à zéro. Cependant, le désavantage est la perte de la propriété de parcimonie : l'estimateur Ridge n'est pas parcimonieux. Une alternative, introduite par [Zou and Hastie, 2005] est l'estimateur *Elastic-Net* où la pénalité est définie par une combinaison convexe des normes ℓ_1 et ℓ_2 : pour $\lambda > 0$ et $\alpha \in [0, 1]$,

$$\hat{\beta}_{\lambda,\text{Elastic-Net}} = \arg\min_{\beta \in \mathbb{R}^p} \Big\{ ||Y - X\beta||_{2,n}^2 + \lambda\Big((1-\alpha)|\beta|_1 + \alpha||\beta||_2^2\Big) \Big\}$$

Le paramètre α réalise ainsi un compromis entre le Lasso (cas $\alpha = 0$) et le Ridge (cas $\alpha = 1$) et conserve la plupart des propriétés de ces deux estimateurs. Par exemple, Elastic-Net conserve la propriété de consistance en sélection asymptotique au sens où $|\beta^*|$, n et p tendent tous les

trois vers l'infini. Grâce à la composante ℓ_2 de la pénalité, l'Elastic-Net est plus régulier que le Lasso et permet de gérer correctement la présence de groupes de variables corrélées, et donc de se libérer de la condition CI du Lasso. En effet, alors que le Lasso sélectionne au hasard une variable parmi un groupe de variables corrélées, l'Elastic-Net encourage les effets de groupes : les variables corrélées sont soit toutes sélectionnées, soit toutes non sélectionnées. Enfin, l'Elastic-Net est efficace lorsque p dépasse n, ce qui n'est pas le cas du Lasso qui est limité à la sélection d'au plus $\min(n, p)$ variables.

Nous proposons une comparaison géométrique sur $\mathbb{R}^2$ des pénalités Lasso, Ridge et Elastic-Net. Sur la Figure 2.3, l'espace sur lequel est minimisé le critère Lasso pour un λ fixé est délimité par le losange en pointillé, celui de l'équation Ridge est délimité par le cercle en pointillé, enfin celui de l'équation Elastic-Net est délimité par la frontière en trait plein. Sans contrainte, la valeur minimale des moindres carrés est représentée par le point A. Sous contrainte, la valeur minimale des moindres carrés pénalisés est le point le plus proche de A à l'intérieur du losange en pointillé, du cercle en pointillé et de la frontière en trait plein pour respectivement le Lasso, le Ridge et l'Elastic-Net. Nous pouvons observer que l'estimateur Lasso est unique et se trouve sur la pointe supérieure du losange en pointillés (correspond à la projection orthogonale du point A sur le losange), ainsi sa deuxième composante est nulle ; l'estimateur Ridge est représenté par la croix violette, aucune de ses coordonnées est nulle. La régularité du cercle en pointillés et son absence de coin rend rare la présence de coordonnées nulles sur l'estimateur (en fait cela se réalise seulement si le point A est sur l'un des axes de coordonnées). Enfin, la frontière d'Elastic-Net est coincée entre le cercle et le losange et est un mélange entre la régularité du cercle et la présence de coin chez le losange.

Enfin, il existe d'autres formes de pénalités utilisées dans la littérature pour obtenir un chemin de régularisation. Nous pouvons citer l'estimateur de Dantzig [Candes et al., 2007], le Fused Lasso [Tibshirani et al., 2005] ou encore l'estimateur du Group-Lasso [Yuan and Lin, 2006] qui proposent de reconsidérer le problème en imposant une parcimonie de groupe : si une variable est sélectionnée, alors le sont aussi toutes celles de son groupe. Cette dernière pénalité est adaptée lorsque les variables sont fortement corrélées mais nécessite la connaissance des groupes a priori. Si ces derniers ne sont pas connus, alors l'Elastic-Net est plus approprié.

- Les algorithmes :

A cause de l'irrégularité de certaines pénalités (notamment la non-dérivabilité), les estimateurs ne sont pas tous accessibles sous une forme explicite. Cependant, la convexité permet de mettre en place des procédures algorithmiques itératives pour approcher numériquement les estimateurs. Nous citons ici deux d'entre eux : l'algorithme *LARS* (Least-angle regression Stagewise) [Efron et al., 2004] et l'algorithme de *descente de gradient* [Friedman et al., 2010]. Ces deux algorithmes sont adaptées aux pénalités Lasso et Elastic-Net qui sont celles utilisées dans cette thèse. L'algorithme LARS repose sur la propriété de linéarité par morceaux de la fonction $\left[\lambda \geq 0 \mapsto \hat{\beta}_\lambda\right]$ pour obtenir le chemin de régularisation : chaque λ considéré correspond à l'entrée d'une variable dans le support. Dans le cadre du Lasso, les valeurs sont explicitées grâce aux *conditions de Karush-Kuhn-Tucker* [Zou et al., 2007]. Il permet de fournir une col-

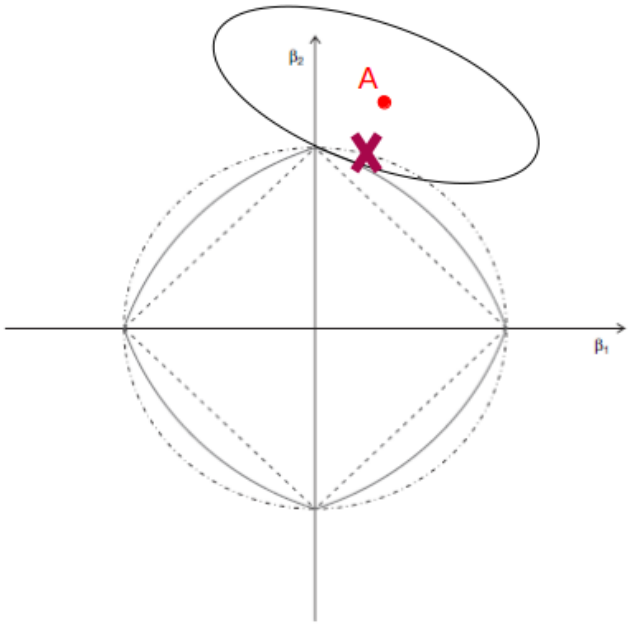


Figure 2.3: Illustration dans $\mathbb{R}^2$ des zones de minimisation des moindres carrés pour les pénalités Lasso (le losange en pointillés), Ridge (Le cercle en pointillés) et Elastic-Net (la frontière en trait plein). Le point rouge A représente la valeur minimale des moindres carrés, la croix violette représente l'estimateur Ridge.

lection de sous-ensembles de variables emboîtés et prend en compte la présence de corrélation entre les variables mais est très sensible au bruit. L'algorithme de descente de gradient consiste à approcher le minimiseur du problème d'optimisation en mettant à jour successivement et coordonnée par coordonnée les coefficients du vecteur solution par descente de gradient. Il est peu sensible au bruit mais les coefficients en sortie sont soit nuls soit significativement non nuls.

- La collection de modèles :
Le chemin de régularisation obtenu est un ensemble de sous-ensembles de variables. Un modèle m est défini par un sous-espace vectoriel de $\mathbb{R}^p$ engendré par certaines des variables $(X_j)_{j\in\{1,\cdots,p\}}$:

$$m = \mathrm{Vect}\Big((X_j)_{j\in J},\ J \in \mathcal{P}(\{1,\cdots,p\})\Big),$$

où $\mathcal{P}(\{1,\cdots,p\})$ désigne toutes les sous-parties de $\{1,\cdots,p\}$. Par exemple, le support de β^* est m^* et il correspond à $\mathrm{Vect}\Big((X_j),\ j \text{ telle que } \beta_j^* \neq 0\Big)$. Nous notons par D_m la dimension du modèle m. Ainsi, chaque sous-ensemble de variables du chemin de régularisation est associé à un modèle et dans cette thèse, nous traitons la question de la sélection du paramètre de régularisation λ comme un problème de sélection de modèles. Au final, quelle que soit la régu-

larisation choisie sur les moindres carrés, la minimisation du critère pénalisé sur une grille finie de λ génère une collection de modèles. Comme celle-ci dépend des données, nous la dénoterons $\mathcal{M}(\mathcal{D})$. Par la suite, nous porterons une attention particulière aux notations $\mathcal{M}(\mathcal{D})$ et $\mathcal{M}$: $\mathcal{M}(\mathcal{D})$ désigne une collection de modèles aléatoire car construite à partir du jeu de données $\mathcal{D}$ disponible, tandis que $\mathcal{M}$ désigne une collection de modèle fixe et déterministe.

- La collection d'estimateurs :
Générer le chemin de régularisation par l'un des algorithmes cités précédemment fournit une collection de modèles mais aussi une collection d'estimateurs puisque les coefficients non-nuls sont estimés dans la procédure. Généralement, et comme mentionné dans le cas du Lasso, ces estimateurs n'ont pas de bonnes propriétés (ils sont biaisés, non-consistants...) : ils minimisent les moindres carrés régularisés et non les moindres carrés. En conséquence, ces estimateurs peuvent être sous-optimaux pour l'estimation du risque et les propriétés oracles ne peuvent pas être satisfaites par les estimateurs Lasso [Fan and Li, 2001]. Choisir l'un d'entre eux n'est donc pas pertinent. Une alternative, mentionnée par [Efron et al., 2004] et explorée par [Connault, 2011], est d'utiliser les algorithmes uniquement pour obtenir la collection de modèle et de ré-estimer ensuite β^* par l'estimateur des moindres carrés sur chaque modèle de la collection :

$$\forall m \in \mathcal{M}(\mathcal{D}), \quad \hat{\beta}_m = \underset{\{\beta, X\beta \in m\}}{\arg\min} ||Y - X\beta||_{2,n}^2. \tag{2.2}$$

Dans cette thèse, nous adopterons cette stratégie : la collection de modèle est $\mathcal{M}(\mathcal{D})$ (ou $\mathcal{M}$ dans le cas d'une collection déterministe) et la collection d'estimateurs est $\left(\hat{\beta}_m\right)_{m \in \mathcal{M}(\mathcal{D})}$ (ou $\left(\hat{\beta}_m\right)_{m \in \mathcal{M}}$ dans le cas d'une collection déterministe). Chaque estimateur des moindres carrés $\hat{\beta}_m$ correspond à un compromis entre ajustement du modèle sur les données et respect du nombre de variables sélectionnées. La sous-section suivante (2.2.2) est dévolue à la sélection de l'estimateur réalisant le meilleur compromis, ceci par minimisation du critère des moindres carrés pénalisés par une fonction ℓ_0.

- Rejouer la construction de la collection de modèles :
L'un des problèmes soulevés quant à la construction du chemin de régularisation est qu'elle est très peu stable [Bach, 2008]. En effet, une légère modification des données (par exemple, l'ajout ou la suppression d'une expérience) peut modifier grandement la collection de modèles obtenue. Pour déjouer cette instabilité et comme un seul jeu de données est disponible en pratique, une idée est de rejouer plusieurs fois la construction du chemin de régularisation à partir de réplicats du jeu de données original. Cela permet de prendre en compte la variabilité entre des jeux de données semblables en plus de la variabilité au sein d'un jeu de données. L'erreur d'estimation due à l'aléa sur les données est ainsi réduite. La première méthode est la *validation-croisée* [Refaeilzadeh et al., 2009] mais celle-ci fournit des résultats pauvres en grande dimension. Une alternative est l'*ESCV* (Estimation Stability with Cross Validation) proposée par [Lim and Yu, 2016] qui consiste à séparer le jeu de données en V groupes. Chacun à leur tour, les groupes constituent le jeu de validation pendant que tous les autres constituent le jeu d'entraînement. Sur chacun de ces derniers est généré une collection de modèles.

Le jeu de validation permet de construire un estimateur qui évalue la stabilité des modèles obtenus (au lieu de servir pour l'évaluation de l'erreur comme la classique validation-croisée V-fold). Parallèlement, [Meinshausen and Bühlmann, 2010] propose des réplicats en créant des sous-échantillons de taille $\lfloor \frac{n}{2} \rfloor$ qui sont obtenus par tirage uniforme et sans remise de $\lfloor \frac{n}{2} \rfloor$ expériences parmi les n du jeu de donnée d'origine. [Bach, 2008] propose, en se basant sur la méthode du *Bootstrap* crée par [Efron and Tibshirani, 1994], des réplicats en créant des ré-échantillons de taille n qui sont obtenus par tirage uniforme et avec remise de n observations parmi les n d'origine. Ces deux méthodes s'appellent respectivement *Stability Selection* et *Bolasso* (pour BOotstrap-enhanced Least Absolute Shrinkage Operator). Pour les deux, un chemin de régularisation est ensuite généré sur chaque réplicat par l'un des algorithmes présentés précédemment. Ces méthodes permettent de distinguer les variables actives des autres. En effet, il est attendu que celles-ci apparaissent un certain nombre de fois dans la plupart des collections. Au contraire, une variable non explicative peut apparaître souvent dans une collection du fait de l'aléa du jeu de données mais est écartée des collections sur la plupart des autres réplicats. En utilisant l'intégralité des modèles des collections, l'ensemble des variables à forte fréquence d'apparition est plus proche de l'ensemble des variables actives. De plus, ce sous-ensemble ne dépend plus a priori de l'aléa du jeu de données initial (l'aléatoire est déplacée sur la construction des réplicats mais comme ceux-ci sont nombreux, son effet est moindre sur l'estimation finale) et reste stable pour n'importe quel échantillon de même taille et réalisé dans les mêmes conditions que l'original. Enfin, un avantage crucial de faire intervenir plusieurs collection de modèles plutôt qu'une seule est la diminution de la variabilité des modèles au sein d'une collection, ce qui réduit l'impact du choix de λ. Certaines propriétés théoriques sur un seul chemin de régularisation sont améliorées. Par exemple, les estimateurs Stability Selection et Bolasso sont consistants en sélection asymptotiquement même si CI n'est pas vérifiée [Zou, 2006, Van de Geer et al., 2011, Bach, 2008]. De plus, la probabilité de sélectionner des variables à tort est contrôlée non-asymptotiquement [Bühlmann and Van De Geer, 2011]. Enfin, des variantes à Stability Selection et Bolasso ont été proposées pour améliorer les performances de l'estimation. Par exemple, [Meinshausen and Bühlmann, 2010] propose de remplacer la pénalité Lasso par la pénalité $\sum_{k=1}^{p} \frac{|\beta_k|}{W_k}$ où chaque W_k est choisi aléatoirement dans $[\alpha, 1]$ pour $\alpha \in]0, 1]$. Ces poids varient selon les réplicats permettant d'écarter des données qui pourraient être des outliers. L'estimateur sous-jacent est appelé le *Randomized Lasso*. [Meinshausen and Bühlmann, 2010] propose également le *Sample Splitting* qui consiste à rajouter systématiquement le complémentaires de chaque réplicat considéré.

2.2.2 La sélection de modèle

La sous-section précédente permet de restreindre l'estimation de $\beta^* \in \mathbb{R}^p$ à une collection de seulement quelques estimateurs pertinents $\left(\hat{\beta}_m\right)_{m \in \mathcal{M}(\mathcal{D})}$ (2.2). Chacun d'eux correspond à un sous-ensemble de variables parmi $\left(X_1, \cdots, X_p\right)$. La dernière étape consiste à sélectionner le sous-ensemble de variables le plus pertinent pour expliquer la variable réponse Y parmi les

candidats pré-sélectionnés. De manière équivalente, cette étape n'est rien d'autre que la calibration du paramètre inconnu intervenant dans la minimisation des moindres carrés régularisés qui donne la collection de modèles (critère Lasso, Elastic-Net, etc...).

Les premières approches pour déterminer le meilleur estimateur de la collection sont la validation-croisée [Refaeilzadeh et al., 2009] qui utilise plusieurs sous-jeux de données et la comparaison directe des maximums de log-vraisemblance des différents modèles. Un résumé des procédures de la validation-croisée est proposée dans [Arlot and Celisse, 2010]. Mais ces méthodes sont longues, coûteuses computationnellement et le support sélectionné n'est pas forcément pertinent. Une alternative est d'évaluer les estimateurs sur leur qualité de prédiction. Dans cette thèse, nous nous intéressons à la procédure de sélection de variables via la sélection de modèle. Elle a été développée pour une collection de modèles fixe $\mathcal{M}$ et sous un point de vue prédictif : l'estimateur final $\hat{\beta}$, construit sur les observations $(y_1, \cdots, y_n, x_{1,1}, \cdots, x_{1,p}, \cdots, x_{n,1}, \cdots, x_{n,p})$, doit permettre de prédire correctement une nouvelle valeur y_{n+1} à partir de nouvelles observations associées $\big(x_{n+1,1}, \cdots, x_{n+1,p}\big)$.

- L'oracle :
Pour un $\beta \in \mathbb{R}^p$ quelconque, l'erreur de prédiction associée est $\mathbb{E}_Y[||Y - X\beta||_{2,n}^2]$. La plus petite erreur de prédiction sur le jeu de données est celle obtenue avec le vrai paramètre β^* : $\mathbb{E}_Y[||Y - X\beta^*||_{2,n}^2]$. C'est pourquoi, la fonction de coût utilisée dans la littérature est l'excès de risque défini par :

$$\forall \beta \in \mathbb{R}^p, \quad \ell(\beta^*, \beta) = \mathbb{E}_Y[||Y - X\beta||_{2,n}^2] - \mathbb{E}_Y[||Y - X\beta^*||_{2,n}^2].$$

Cette fonction de perte est en effet adaptée au problème d'estimation de β^* puisque le minimum est atteint en $\beta = \beta^*$, et ce quel que soit le jeu de données (Y, X) disponible. De plus, en utilisant que $\mathbb{E}_Y[\varepsilon] = 0$, on obtient :

$$\forall \beta \in \mathbb{R}^p, \quad \ell(\beta^*, \beta) = \mathbb{E}_Y[||X\beta - X\beta^*||_{2,n}^2]$$

En effet,

$$\begin{aligned}
\forall \beta \in \mathbb{R}^p, \quad \ell(\beta^*, \beta) &= \mathbb{E}_Y[||Y - X\beta||_{2,n}^2] - \mathbb{E}_Y[||Y - X\beta^*||_{2,n}^2] \\
&= \mathbb{E}_Y[\cancel{||Y||_{2,n}^2} + ||X\beta||_{2,n}^2 - \frac{2}{n}\langle Y, X\beta\rangle - \cancel{||Y||_{2,n}^2} - ||X\beta^*||_{2,n}^2 + \frac{2}{n}\langle Y, X\beta^*\rangle] \\
&= \mathbb{E}_Y[||X\beta||_{2,n}^2 - \frac{2}{n}\langle X\beta^* + \varepsilon, X\beta\rangle - ||X\beta^*||_{2,n}^2 + \frac{2}{n}\langle X\beta^* + \varepsilon, X\beta^*\rangle] \\
&= \mathbb{E}_Y[||X\beta||_{2,n}^2 - \frac{2}{n}\langle X\beta^* + \varepsilon, X\beta\rangle + ||X\beta^*||_{2,n}^2 + \frac{2}{n}\langle \varepsilon, X\beta^*\rangle] \\
&= \mathbb{E}_Y[||X\beta||_{2,n}^2 - \frac{2}{n}\langle X\beta^*, X\beta\rangle + ||X\beta^*||_{2,n}^2 + \frac{2}{n}\langle \varepsilon, X\beta^* - X\beta\rangle] \\
&= \mathbb{E}_Y[||X\beta||_{2,n}^2 - \frac{2}{n}\langle X\beta^*, X\beta\rangle + ||X\beta^*||_{2,n}^2] + \cancel{\frac{2}{n}\mathbb{E}_Y[\langle \varepsilon, X\beta^* - X\beta\rangle]} \\
&= \mathbb{E}_Y[||X\beta - X\beta^*||_{2,n}^2]. \qquad (2.3)
\end{aligned}$$

Ainsi, le meilleur estimateur d'un point de vue prédictif parmi les estimateurs des moindres carrés $\left(\hat{\beta}_m\right)_{m\in\mathcal{M}}$ disponibles dans la collection $(m)_{m\in\mathcal{M}}$ est obtenu en minimisant le risque prédictif défini en (2.3).

Introduisons le modèle oracle m_0 et son risque r_{m_0}, appelé risque oracle :

$$m_0 = \arg\min_{m\in\mathcal{M}} \mathbb{E}_Y[||X\hat{\beta}_m - X\beta^*||_{2,n}^2]$$

$$r_{m_0} = \inf_{m\in\mathcal{M}} \mathbb{E}_Y[||X\hat{\beta}_m - X\beta^*||_{2,n}^2] = \mathbb{E}_Y[||X\hat{\beta}_{m_0} - X\beta^*||_{2,n}^2].$$

La quantité r_{m_0} est le plus petit risque prédictif possible sur la collection de modèles ; ainsi, m_0 est le meilleur modèle de $\mathcal{M}$ d'un point de vue prédictif. Notons que même si m^* est dans la collection de modèles, m^* n'est pas nécessairement égal à m_0.
En effet,

$$X\beta_{m^*} = \Pi_{m^*}(X\beta^*) = X\beta^*$$

mais

$$X\hat{\beta}_{m^*} = \Pi_{m^*}(Y)$$

peut être différent de $X\beta^*$. Or, $X\hat{\beta}_{m_0}$ est proche de $X\beta^*$ en norme ℓ_2, donc $X\hat{\beta}_{m^*}$ n'est pas nécessairement proche de $X\hat{\beta}_{m_0}$ en norme ℓ_2 et ne sera pas le meilleur modèle pour la prédiction. Ainsi, en sélection de modèle, le modèle référent est m_0 et non pas m^*. Comme celui-ci dépend du paramètre inconnu β^*, m_0 est incalculable à partir du jeu de données et le challenge à relever est alors le suivant : comment sélectionner $\hat{m}$ uniquement à partir du jeu de données de telle sorte que soit garantie d'un point de vue théorique que le risque prédictif associé à $\hat{m}$ soit le plus proche possible du risque oracle r_{m_0} ?

- Minimisation d'un contraste pénalisé :
Minimiser le risque prédictif (2.3) est impossible en pratique à cause de sa dépendance en β^*, le paramètre inconnu de la régression. En pratique, il est approché par une fonction contraste empirique notée γ_n qui est évaluée sur le jeu de données. Dans le cadre de la régression linéaire, les fonctions populaires sur $\beta \in \mathbb{R}^p$ sont $-\log(\text{Vrai}(Y;(\beta,\hat{\sigma}^2)))$ où $\log(\text{Vrai})$ désigne la log-vraisemblance du modèle avec $\hat{\sigma}^2$ un estimateur de la variance σ^2, et les moindres carrés empiriques définis ci-dessous :

$$\forall\beta \in \mathbb{R}^p, \quad \gamma_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2. \tag{2.4}$$

Cependant, comme expliqué précédemment, maximiser la log-vraisemblance ou minimiser (2.4) en $\beta \in \mathbb{R}^p$ n'est pas adapté en grande-dimension où l'hypothèse de parcimonie est nécessaire : les estimateurs sous-jacents sont pleins et perdent leurs propriétés (sans biais, etc...). C'est pourquoi, la procédure de sélection de modèles repose sur la minimisation en $m \in \mathcal{M}$ du critère suivant :

$$\text{crit}(m) = \gamma_n(\hat{\beta}_m) + \text{pen}(m), \tag{2.5}$$

où pen est une fonction positive, croissante et dépendante de D_m. Le modèle final sélectionné $\hat{m}$ est défini par :

$$\hat{m} = \arg\min_{m \in \mathcal{M}} \Big\{ \text{crit}(m) \Big\}, \tag{2.6}$$

il réalise un compromis entre l'ajustement des données au modèle (à travers $\gamma_n(\hat{\beta}_m)$) et le respect de la parcimonie (à travers pen(m)). Tout le challenge réside en le choix d'une bonne fonction de pénalité pen de sorte que le modèle sélectionné associé $\hat{m}$ réalise un risque prédictif le plus proche possible de r_{m_0}.

Dans le cas gaussien, maximiser la log-vraisemblance est équivalent à minimiser les moindres carrés : en effet, ces deux quantités sont égales à constantes multiplicative et additive près. Cependant, la log-vraisemblance fait intervenir la variance σ^2. Lorsque celle-ci est inconnue, elle est en général estimée par son estimation empirique classique à moyenne connue :

$$\frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} x_{ij} \hat{\beta}_{\lambda,j} \Big)^2$$

rendant l'évaluation de la log-vraisemblance possible mais asymptotique en pratique. Nous verrons par la suite qu'il est préférable de choisir les moindres carrés (indépendant de σ^2) pour le contraste et de faire intervenir σ^2 dans la pénalité puisque des méthodes existent pour gérer le cas de la variance inconnue au sein de la fonction de pénalité. De plus, en valeur absolue, les valeurs des moindres carrés sont plus petites que celles de la log-vraisemblance qui peuvent être très grandes, ce qui peut engendrer des problèmes computationnels.

- Les pénalités asymptotiques :
Historiquement, [Akaike, 1973] et [Schwarz et al., 1978] sont les premiers à avoir proposé des fonctions de pénalités avec comme contraste la log-vraisemblance. Les estimateurs $\hat{\beta}_{\hat{m}}$ correspondant satisfont une égalité oracle asymptotique sur le risque prédictif :

$$\frac{\mathbb{E}_Y[||X\hat{\beta}_{\hat{m}} - X\beta^*||_{2,n}^2]}{r_{m_0}} \underset{n \to +\infty}{\longrightarrow} 1 \tag{2.7}$$

Tous les deux supposent une collection de modèles fixée et se sont basés sur des approximations asymptotiques. Notons que Akaike propose, par le théorème de Wilks, le critère *Akaike Information Criterium* (*AIC* [Akaike, 1973]) avec

$$\text{pen}_{\text{AIC}}(m) = D_m;$$

alors que G.Schwarz propose, par une approche bayésienne, le *critère Bayesian Information Criterium* (*BIC* [Schwarz et al., 1978]) avec

$$\text{pen}_{\text{BIC}}(m) = \frac{D_m \log(n)}{2}.$$

Ainsi, par cette approche bayésienne, une loi a priori est choisie sur les modèles. Le terme $\log(n)$ rend la procédure BIC plus conservative que AIC. Au même moment, [Mallows, 2000]

étudie ces critères pénalisés dans le cadre de la régression et propose, avec les moindres carrés comme contraste,

$$\mathrm{pen}_{\mathrm{Mallows}}(m) = \frac{2D_m\sigma^2}{n}.$$

Tout comme AIC et BIC, le critère *CP.Mallows* ne prend pas en compte la variabilité du terme $||Y - X\hat{\beta}_m||_{2,n}^2$ autour de $||X\hat{\beta} - X\beta^*||$, ce qui est un problème lorsque le nombre de modèles pour une dimension fixée croît exponentiellement en la dimension (ce qui est notre cas) : l'estimateur sélectionné par AIC, BIC ou Mallows a tendance à avoir un support très grand et fait face à des problèmes de sur-apprentissage. Une alternative a été proposée par [Chen and Chen, 2008] et inspirée de BIC : alors que G.Schwarz choisit la loi uniforme pour ne privilégier aucun modèle, [Chen and Chen, 2008] propose une autre loi a priori faisant intervenir le nombre de modèle par dimension : $\log\left(\binom{p}{D_m}\right)$, ce qui donne

$$\mathrm{pen}_{\mathrm{eBIC}}(m) = \frac{D_m \log(n)}{2} + \frac{\gamma}{2}\log\left(\binom{p}{D_m}^{-1}\right),$$

avec $\gamma \in [0,1]$. Cette pénalité est donc plus forte que les précédentes, ainsi, la procédure *eBIC* (pour *extended Bayesian Information Criteriua*) est plus conservative. Construits via des chemins différents, ces quatre critères n'ont pas les mêmes propriétés asymptotiques [Yang, 2005] : AIC et Mallows offrent des estimateurs sans biais du risque ; si le vrai modèle m^* appartient à la collection, BIC et eBIC assurent que la probabilité de le choisir tend vers 1 quand n tend vers l'infini (estimateurs consistants) ; eBIC assure en plus que, si p est raisonnable (au plus une puissance de n), la probabilité de sélectionner un modèle différent de m^* tend vers 0. De plus, [Yang, 2005] a montré que satisfaire toutes ces propriétés simultanément est impossible : BIC ne respecte pas la propriété d'optimalité asymptotique ; l'estimateur d'AIC n'est pas consistant, etc... Ces critères aux propriétés asymptotiques sont pauvres d'un point de vue prédictif pour un contexte de grande dimension [Giraud, 2014, Baraud et al., 2009]. Des propriétés vérifiées pour toutes valeurs de n et p fixées sont préférables.

- Les pénalités non-asymptotiques :

Une approche non-asymptotique est préférable : les propriétés sur les estimateurs sont vraies pour toutes valeurs de n et p. Dans cette direction, les auteurs de [Birgé and Massart, 2001] ont mis en place une procédure pour sélectionner $\hat{m}$ telle que, quel que soit le jeu de données (Y, X) et la collection d'estimateurs linéaires disponible, $\hat{\beta}_{\hat{m}}$ satisfasse l'inégalité oracle suivante :

$$\mathbb{E}_Y[||X\hat{\beta}_{\hat{m}} - X\beta^*||_{2,n}^2] \leq C_n r_{m_0} + R_n, \tag{2.8}$$

où $C_n \approx 1$ au moins pour n grand et R_n est petit comparé à r_{m_0}. Notons qu'un estimateur vérifiant (2.8) vérifie (2.7) [Arlot, 2011].

Pour cela, ce mêmes auteurs définissent dans [Birgé and Massart, 2007] successivement une fonction de pénalité idéale, une fonction de pénalité minimale et une fonction de pénalité optimale. Nous définissons ci-dessous chacune d'entre elles.

- *La fonction de pénalité idéale.*

Observons qu'à partir de la définition de $\hat{m}$ (2.6), nous obtenons :

$$\gamma_n(\hat{\beta}_{\hat{m}}) + \text{pen}(\hat{m}) = \inf_{m\in\mathcal{M}}\Big\{\gamma_n(\hat{\beta}_m) + \text{pen}(m)\Big\}.$$

La fonction de pénalité idéale est donc $\text{pen} = \text{pen}^*_{\text{ideal}}$ telle que :

$$\text{pen}^*_{\text{ideal}}(m) = \mathbb{E}_Y[||X\hat{\beta}_m - X\beta^*||^2_{2,n}] - \gamma_n(\hat{\beta}_m). \tag{2.9}$$

Malheureusement, $\text{pen}^*_{\text{ideal}}$ dépend de β^* mais on obtient avec la définition de $\text{pen}^*_{\text{ideal}}$ (2.9) :

$$\mathbb{E}_Y[||X\hat{\beta}_{\hat{m}}-X\beta^*||^2_{2,n}]+[\text{pen}(\hat{m})-\text{pen}^*_{\text{ideal}}(\hat{m})] = \inf_{m\in\mathcal{M}}\Big\{\mathbb{E}_Y[||X\hat{\beta}_m-X\beta^*||^2_{2,n}]+[\text{pen}(m)-\text{pen}^*_{\text{ideal}}(m)]\Big\}. \tag{2.10}$$

Finalement, pour obtenir une inégalité oracle comme (2.8), la fonction de pénalité pen disponible sur le jeu de données doit être proche de $\text{pen}^*_{\text{ideal}}$, et même la fonction $\text{pen} - \text{pen}^*_{\text{ideal}}$ doit être bornée supérieurement (pour le terme de droite dans (2.10)) et inférieurement (pour le terme de gauche dans (2.10)). Dans (2.9), seul le terme $\mathbb{E}_Y[||X\hat{\beta}_m - X\beta^*||^2_{2,n}]$ est indisponible sur les données. Pour pallier à ce problème, L.Birgé et P.Massart utilise la décomposition suivante du risque prédictif : pour tout $m \in \mathcal{M}$,

$$\begin{aligned}\mathbb{E}_Y[||X\hat{\beta}_m - X\beta^*||^2_{2,n}] = \mathbb{E}_Y[||X\beta_m - X\beta^*||^2_{2,n}] + \mathbb{E}_Y[||X\hat{\beta}_m - X\beta_m||^2_{2,n}]\\ + \frac{2}{n}\mathbb{E}_Y[\langle X\hat{\beta}_m - X\beta_m, X\beta_m - X\beta^*\rangle]\end{aligned} \tag{2.11}$$

Pour chaque $m \in \mathcal{M}$, $X\beta_m = \Pi_m(X\beta^*)$, donc $X\beta_m - X\beta^*$ appartient à $m^\perp$ et donc $\mathbb{E}_Y[\langle X\hat{\beta}_m - X\beta_m, X\beta_m - X\beta^*\rangle] = 0$. De plus, $||X\beta_m - X\beta^*||^2_{2,n}$ est déterministe. Donc, l'équation (2.11) devient:

$$\begin{aligned}\forall m \in \mathcal{M}, \qquad \mathbb{E}_Y[||X\hat{\beta}_m - X\beta^*||^2_{2,n}] &= ||X\beta_m - X\beta^*||^2_{2,n} + \mathbb{E}_Y[||X\hat{\beta}_m - X\beta_m||^2_{2,n}]\\ &\underset{(*)}{=} ||(I_n - \Pi_m)(X\beta^*)||^2_{2,n} + \sigma^2\frac{D_m}{n}.\end{aligned} \tag{2.12}$$

La clé de cette décomposition est que les estimateurs $\hat{\beta}_m$ sont les projections linéaires de Y sur les modèles m. Ainsi, *(*)* est vraie puisque $X\hat{\beta}_m - X\beta_m = \Pi_m(Y) - \Pi_m(X\beta^*) = \Pi_m(Y - X\beta^*) = \Pi_m(\varepsilon)$; et donc comme $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, on obtient par application du *théorème de Cochran* que $\mathbb{E}_Y[||\Pi_m(\varepsilon)||^2_{2,n}] = \sigma^2\frac{D_m}{n}$.

Cette décomposition (2.12) est appelée *décomposition biais-variance*. Le premier terme, correspondant au terme de biais, est déterministe et reflète la qualité du modèle m pour donner une estimation de $X\beta^*$. Le second terme, correspondant au terme de la variance, et dépend de la variabilité du jeu de données disponible. Le terme de biais décroît en fonction de la dimension du modèle tandis que le terme de variance croît. L'oracle m_0 atteint le meilleur compromis entre ces deux erreurs.

- *La fonction de pénalité minimale.*

En appliquant la théorie des grandes déviations et des inégalités de concentration sur le terme de bais, L.Birgé and P.Massart obtiennent dans [Birgé and Massart, 2001], pour une variance σ^2 connue et une collection de modèles exhaustive, une borne inférieure sur pen pour obtenir une inégalité oracle. Grâce à des inégalités plus fines, ils définissent ensuite dans [Birgé and Massart, 2007] une fonction de pénalité minimale donnée par une forme explicite de telle sorte que toutes fonctions supérieures à celle-ci satisfassent une inégalité oracle :

Théorème 2.1. *([Birgé and Massart, 2007]).*
Soit $\mathcal{M}$ la collection de modèles déterministe et $\{L_m\}_{m\in\mathcal{M}}$ une famille de poids, i.e. une famille de nombres réels positifs ou nuls, satisfaisant :

$$\Sigma = \sum_{\{m\in\mathcal{M}, D_m>0\}} e^{-L_m D_m} < +\infty \tag{2.13}$$

Soit deux nombres $\theta \in]0,1[$ et $\kappa > 2-\theta$. Supposons qu'il existe une famille $\overline{\mathcal{M}}$ (possiblement finie) de sous-ensembles de $\mathcal{M}$ telle que la fonction de pénalité satisfait : $\forall m \in \mathcal{M}\backslash\overline{\mathcal{M}}$

$$pen(m) \geq \sigma^2 \frac{D_m}{n}\Big(\kappa + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m\Big), \tag{2.14}$$

alors, l'estimateur de projection associé $\hat{\beta}_{\hat{m}}$ existe presque sûrement et est unique. De plus, il vérifie une inégalité oracle non-asymptotique sur le risque prédictif :

$$\begin{aligned}(1-\theta)\ \mathbb{E}[||X\hat{\beta}_{\hat{m}} - X\beta^*||_{2,n}^2] \leq \inf_{m\in\mathcal{M}}\Big\{ &\inf_{\{\beta\in\mathbb{R}^p, X\beta\in m\}}\{||X\beta - X\beta^*||_{2,n}^2\} + pen(m) - \sigma^2\frac{D_m}{n}\Big\} \\ &+ \sup_{m\in\overline{\mathcal{M}}}\Big\{Q(m) - pen(m)\Big\} + \sigma^2\Sigma\Big((2-\theta)^2(\kappa+\theta-2)^{-1} + 2\theta^{-1}\Big)\end{aligned} \tag{2.15}$$

où

$$\forall m \in \mathcal{M}, \quad Q(m) = \sigma^2\frac{D_m}{n}\Big(\kappa + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m\Big).$$

Ainsi, toute pénalité plus grande que Q_m sur $\mathcal{M}$ offre une inégalité oracle. De plus, la contribution de [Birgé and Massart, 2007] par rapport à [Birgé and Massart, 2001] est qu'une forme explicite pour pen est désormais accessible à travers l'utilisation de Q.
La fonction de pénalité minimale est définie de telle sorte que, pour toute fonction de pénalité supérieure à cette pénalité minimale, une inégalité oracle du type (2.8) est satisfaite. Ainsi, d'après le Théorème 2.1, la fonction de pénalité minimale appropriée est :

$$\forall m \in \mathcal{M}, \quad \text{pen}^*_{\min}(m) = \sigma^2\frac{D_m}{n}\Big(\kappa + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m\Big), \quad \text{for } \kappa > 2-\theta \quad \text{and } \theta \in]0,1[. \tag{2.16}$$

- *La fonction de pénalité optimale.*

D'après le Théorème 2.1, toute fonctions de pénalité supérieures à $\text{pen}^*_{\text{min}}$ offrent une inégalité oracle. Comme la pénalité minimale est connue de manière explicite, la forme la plus simple pour une pénalité optimale est $K\text{pen}^*_{\text{min}}$ pour $K > 1$. Ainsi, la fonction de pénalité optimale est définie par : $\forall m \in \mathcal{M},\ \forall \theta \in]0,1[,\ \forall \kappa > 2-\theta,\ \forall K > 1$,

$$\text{pen}^*_{\text{opt}}(m) = K\sigma^2 \frac{D_m}{n}\Big(\kappa + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m\Big). \tag{2.17}$$

Cette pénalité dépend de n, p et σ^2 et prend en argument $(m, D_m)_{m\in\mathcal{M}}$. En pratique, elle ne dépend de la collection $(m)_{m\in\mathcal{M}}$ qu'à travers la dimension des modèles $(D_m)_{m\in\mathcal{M}}$ (L_m ne fera intervenir que les Dm).

Trois problèmes méritent d'être soulevés :

1. Comment choisir la valeur de $K > 1$? Comme la constante K apparaît dans la borne supérieure du risque prédictif (2.15) à travers pen, une valeur trop grande n'est pas souhaitable. En pratique, $K = 2$ est souvent utilisée car 2 est une constante optimale asymptotique. Nous renvoyons au chapitre 5 pour plus de détails. Par ailleurs, des résultats théoriques montrent que 2 est une constante toujours raisonnable (sans pour autant être optimale) d'un point de vue non-asymptotique [Birgé and Massart, 2007]. Nous proposons une calibration de K dans le Chapitre 5.

2. Comment déterminer en pratique les constantes θ et κ ?

3. Comment gérer le cas de la variance σ^2 inconnue ?

Les deux derniers item sont traités dans le chapitre 5 où une calibration des deux constantes est proposée, celle-ci englobe la gestion de la variance inconnue. Concernant le dernier item, plusieurs solutions sont proposées pour estimer ce paramètre. Par exemple :

- Choisir un modèle m arbitraire dans la collection et estimer la variance par son estimateur classique au sein de m :
$$\widehat{\sigma^2_m} = \frac{||Y - X\hat{\beta}_m||^2_{2,n}}{n - D_m}$$
Cette procédure dépend beaucoup du choix de m et est peu efficace.

- Pour chaque $m \in \mathcal{M}$, estimer la variance à l'intérieur des sous-espaces correspondant :
$$\forall m \in \mathcal{M},\quad \widehat{\sigma^2_m} = \frac{||Y - X\hat{\beta}_m||^2_{2,n}}{n - D_m}$$

Ces deux méthodes dépendent beaucoup de la collection de modèles disponible et offrent des procédures instables. C'est pourquoi ont vu le jour des procédures estimant σ^2 indépendamment de la collection de modèles. La première idée est l'estimateur empirique mais celui-ci ne peut pas

être utilisé car sa propriété de consistance fait défaut en grande dimension. Dans cette thèse, nous nous intéresserons aux deux critères non-asymptotiques adaptés à notre cadre : grande dimension, variance σ^2 inconnue et large collection de modèles. Les deux sont issus des travaux théoriques de [Birgé and Massart, 2001, Birgé and Massart, 2007] et offrent donc des garanties théoriques non-asymptotiques sur le risque prédictif. Le premier est le critère *LinSelect* [Baraud et al., 2009], le second contient les *pénalités dépendantes des données* [Birgé and Massart, 2007].

- *La pénalité LinSelect.*

Le critère LinSelect a été introduit dans [Baraud et al., 2009] et généralisé au cadre de la grande dimension dans [Giraud et al., 2012]. Pour l'obtenir, nous avons besoin des définitions suivantes :

Définition 2.1. Soient D et$N$ deux nombres positifs et X_D, X_N deux variables aléatoires indépendantes de lois respectives $\chi^2(D)$ et $\chi^2(N)$. Pour $x \geq 0$, on définit :

$$\text{Dkhi}[D, N, x] = \frac{1}{D}\mathbb{E}\Big[\max\big(0, X_D - x\frac{X_N}{N}\big)\Big] \tag{2.18}$$

Définition 2.2. Soient D et$N$ deux entiers positifs. Pour $0 < q \leq 1$, on définit EDkhi[D,N,q] comme l'unique solution de l'équation :

$$\text{Dkhi}[D, N, \text{EDkhi[D,N,q]}] = q \tag{2.19}$$

Le théorème suivant définit la pénalité LinSelect : $\text{pen}_{\text{LinSelect}}$ vérifiant une inégalité oracle :

Théorème 2.2. *([Baraud et al., 2009]).*
Soit $\mathcal{M}(\mathcal{D})$ une collection de modèles aléatoire et $\{L_m\}_{m\in\mathcal{M}(\mathcal{D})}$ une famille de poids vérifiant :

$$\Sigma := \sum_{\{m\in\mathcal{M}(\mathcal{D}),\ D_m>0\}} (D_m + 1)e^{-L_m} < +\infty \tag{2.20}$$

Soit $K > 1$ et soit la pénalité suivante : $\forall m \in \mathcal{M}(\mathcal{D})$,

$$pen_{LinSelect}(m) = \frac{K}{n} \times \frac{n - D_m}{n - D_m - 1} EDkhi\ [D_m + 1, n - D_m - 1, e^{-L_m}]. \tag{2.21}$$

Alors, l'estimateur correspondant sélectionné $\hat{\beta}_{\hat{m}}$ vérifie l'inégalité oracle suivante :

$$\mathbb{E}_Y\Big[\frac{||X\beta^* - X\hat{\beta}_{\hat{m}}||^2_{2,n}}{\sigma^2}\Big] \leq \frac{K}{K-1} \inf_{m\in\mathcal{M}(\mathcal{D})} \Big\{ \frac{||X\beta^* - X\beta_m||^2_{2,n}}{\sigma^2} \times \tag{2.22}$$

$$\Big(1 + \frac{pen_{LinSelect}(m)}{n - D_m}\Big) + pen_{LinSelect}(m) - D_m \Big\}$$

$$+ 2K^2 \frac{\Sigma}{K-1}. \tag{2.23}$$

Ici, la variance est estimée sur chaque modèle m par son estimateur empirique. La constante K a été fixée dans [Giraud et al., 2012] à 1.1 via une large étude de simulations, ce qui rend la pénalité $\text{pen}_{\text{LinSelect}}$ complètement déterministe.

- *Les pénalités data-dépendantes.*

Les deux principales idées des pénalités data-dépendantes est de considérer les constantes inconnues de la pénalité comme des hyperparamètres à calibrer directement sur le jeu de données et d'inclure la variance σ^2 dans ces hyperparamètres.
Partons de la définition de la pénalité minimale (2.16). Les inégalités suivantes permettent d'obtenir la famille de poids $\{L_m\}_{m\in\mathcal{M}(\mathcal{D})}$ adaptée à notre modèle (1.1) et à nos hypothèses (grande dimension et collection de modèles aléatoires) :

$$\begin{aligned}\Sigma = \sum_{\{m\in\mathcal{M}(\mathcal{D}),D_m>0\}} e^{-L_m D_m} &= \sum_{D=1}^{p} e^{-L_D D}\ \text{Card}\Big(\{m\in\mathcal{M}(\mathcal{D}), D_m = D\}\Big)\\ &\underset{(ii)}{\leq} \sum_{D=1}^{p} e^{-L_D D}\binom{p}{D} = \sum_{D=1}^{p} e^{-D\left(L_D - \frac{\log\left(\binom{p}{D}\right)}{D}\right)}\end{aligned}$$

(ii) est atteint par la collection exhaustive complète contenant tous les modèles jusqu'à $D_m = p$. Pour avoir la condition (2.13) du théorème et puisque Σ apparaît dans la borne supérieure de l'inégalité oracle (2.15), nous choisissons L_D telle que $\Sigma < 1$. Cela entraîne que $L_D = \delta + \frac{\log\left(\binom{p}{D}\right)}{D}$ pour n'importe quel $\delta > 0$. Ainsi, la pénalité minimale (2.16) est ré-écrit en : pour tout $\theta \in]0,1[$ et $\kappa > 2-\theta$:

$$\text{pen}^*_{\min}(m) = \sigma^2\kappa\frac{D_m}{n} + \sigma^2 2(2-\theta)\frac{D_m}{n}\sqrt{\delta + \frac{\log(\binom{p}{D_m})}{D_m}} + \frac{2\sigma^2}{\theta}\frac{D_m}{n}\Big(\delta + \frac{\log(\binom{p}{D_m})}{D_m}\Big) \tag{2.24}$$

D'un point de vue applicatif, fixer les constantes $\theta \in]0,1[, \kappa > 2-\theta$ et $\delta > 0$ assure que pour toute fonction de pénalité plus grande que celle obtenue, une inégalité oracle est vérifiée. Cependant, la fonction obtenue n'est pas nécessairement la pénalité minimale théorique (2.16). De plus, notre choix de poids $\Big(L_m\Big)_{m\in\mathcal{M}(\mathcal{D})}$ n'est pas nécessairement le choix optimal. C'est pourquoi, afin de garder une flexibilité dans $\text{pen}^*_{\min}$, les hyperparamètres sont estimés directement sur le jeu de données. Pour réduire le nombre d'entre eux, nous utilisons $2ab \leq a^2 + b^2$ sur le second terme de la somme dans (2.16) avec $a = \sqrt{1}$ et $b = \sqrt{\delta + \frac{\log(\binom{p}{D})}{D}}$. Une pénalité minimale disponible sur un jeu de données est ainsi obtenue :

$$\begin{aligned}&\forall m\in\mathcal{M}(\mathcal{D}), \quad \text{pen}_{\min}(m) = \frac{D_m}{n}\sigma^2\left(\kappa_1 + \kappa_2\frac{\log(\binom{p}{D_m})}{D_m}\right)\\ &\text{ou encore :} \qquad \text{pen}_{\min}(m) = \left(C_1(\sigma^2)\frac{D_m}{n} + C_2(\sigma^2)\frac{\log\left(\binom{p}{D_m}\right)}{n}\right)\end{aligned}$$

où $C_1(\sigma^2) = \sigma^2\kappa_1$ et $C_2(\sigma^2) = \sigma^2\kappa_2$ sont les deux uniques constantes inconnues qui dépendent de σ^2. Par la suite, nous conserverons la notation la plus générale $C_1(\sigma^2)$ et $C_2(\sigma^2)$ car nous n'exploiterons pas les formes $\sigma^2\kappa_1$ et $\sigma^2\kappa_2$.

D'un point de vue pratique, la méthode dite de l'*heuristique de pente* permet d'estimer les hyperparamètres $C_1(\sigma^2)$ et $C_2(\sigma^2)$ via l'algorithme du saut de dimension ou via l'algorithme d'estimation d'une pente. Cette méthode est construite sur des aspects théoriques et sur des idées heuristiques. Elle conduit à la définition de fonctions $\mathrm{pen}_{\mathrm{opt}}$ qui sont dites dépendantes des données. Elle a été introduite dans le cadre de la régression linéaire gaussienne (5.1) avec une variance σ^2 connue, une matrice de design X fixe, une exploration exhaustive possible de tous les modèles et non dans le cadre de la grande dimension. Ses travaux ont ensuite été généralisés au cadre hétéroscédastique [Arlot and Massart, 2009]. Le principe de l'heuristique de pente a également été validé d'un point de vue théorique pour l'estimation de densité sur les moindres carrés dans [Lerasle, 2009] et pour la sélection de voisinage dans un champ aléatoire gaussian et markovien. Plus de détails sur la méthode de l'heuristique de pente sont donnés dans le chapitre 5.

Pour conclure, la procédure de sélection de modèle adaptée à notre modèle (1.1) et à nos hypothèses de grande dimension, de collection de modèles aléatoires et de variance σ^2 inconnue, est appliquée en minimisant (2.5) avec la fonction de pénalité optimale définie par :

$$\mathrm{pen}_{\mathrm{opt}}(m) = K\left(C_1(\sigma^2)\frac{D_m}{n} + C_2(\sigma^2)\frac{\log(\binom{p}{D_m})}{n}\right),$$

où $K > 1$, $C_1(\sigma^2)$ et $C_2(\sigma^2)$ sont les trois uniques constantes inconnues qui dépendent du design expérimental uniquement à travers σ^2. Pour des considérations pratiques, ces constantes sont des hyperparamètres. Notons que l'apparition du terme binomial dans la pénalité est récente pour des considérations pratiques puisque son implémentation nécessitait un coût computationnel important.

2.3 Contrôle du FDR dans un cadre de tests multiples

En régression linéaire gaussienne, une autre méthode pour sélectionner des variables est de tester la pertinence de chacune d'entre elles par rapport à Y. Nous introduisons dans un premier temps la notion de tests puis, dans un second temps, le contexte de grande dimension oblige à considérer l'ensemble de ces p tests simultanément : c'est le principe des tests multiples. Nous définissons les fonctions de coût classiques avant de donner une liste non exhaustive d'autres fonctions de coût étudiées dans ce contexte.

- Tests statistiques :
Soient $(X_1, \cdots, X_p)$ une famille de variables aléatoires de lois respectives $(P_1, \cdots, P_p)$. Ici, le but n'est pas de retrouver les P_i mais de tester si elles appartiennent aux classes de distribution

$\Theta_{0,i}$ ou aux classes de distribution $\Theta_{1,i}$. Nous définissons les *hypothèses nulles* et les *hypothèses alternatives* des tests de la façon suivante :

$$\forall i \in \{1, \cdots, p\}, \quad H_{0,i} : P_i \in \Theta_{0,i} \qquad VS \qquad H_{1,i} : P_i \in \Theta_{1,i}$$

Usuellement, la famille $\Theta_{1,i}$ correspond à des rejets de $\Theta_{0,i}$, alors que $\Theta_{0,i}$ joue le rôle de contrôle : on conserve $H_{0,i}$ si le test de rejet n'a pas été significatif. Les deux hypothèses ne sont donc pas, en général, interchangeables : on préfère se tromper sous $H_{1,i}$, c'est-à-dire rater un rejet, plutôt que sous $H_{0,i}$, c'est-à-dire déclarer une faux rejet. Les procédures de tests ont pour but de déterminer la bonne hypothèse pour chaque $i \in \{1, \cdots, p\}$.

Pour cela, deux méthodes existent. A partir d'une *statistique de test* $\widehat{S}_i$ (fonction de X_i) dont la loi est connue sous $H_{0,i}$ et est différente de celle sous $H_{1,i}$, une zone de rejet $\widehat{R}_i$ est construite. Elle correspond à certaines valeurs extrêmes de la distribution sous $H_{0,i}$. Une observation tombant dans la zone de rejet a donc une faible probabilité d'être une réalisation de la loi sous $H_{0,i}$ et on la considérera comme une réalisation de la distribution $H_{1,i}$. La probabilité de se tromper, c'est-à-dire de rejeter $H_{0,i}$ à tort est faible. Le seuil définissant la zone de rejet de $H_{0,i}$ est ainsi calculé de sorte que la probabilité de rejeter $H_{0,i}$ à tort est contrôlée par α, pour un α donné (on dit que le test est de *niveau* α) ; mais aussi de sorte que le test fasse le plus de rejets possibles. Ainsi, ce seuil est tel que la probabilité de se tromper sous $H_{0,i}$ soit contrôlée par α mais soit la plus proche de α. Le choix du paramètre α est important puisqu'il réalise un compromis entre l'*erreur de première espèce* (la probabilité de se tromper sous $H_{0,i}$, c'est-à-dire déclarer des faux rejets) et l'*erreur de seconde espèce* (la probabilité de se tromper sous $H_{1,i}$, c'est-à-dire rater des rejets). Par exemple, un α trop petit empêche la détection de faux rejets mais de nombreux rejets sont ratés : l'erreur de seconde espèce est grande, alors qu'un α trop grand entraîne trop de rejets : l'erreur de première espèce est grande. Il n'est pas possible de contrôler parfaitement les deux erreurs simultanément. En pratique, $\alpha = 0.05$. Enfin, la statistique de test est évaluée sur un jeu de données : si la valeur observée $\widehat{S}_i(x_{i,\text{obs}})$, où $x_{i,\text{obs}}$ est la valeur observée de X_i, est dans la zone de rejet de $H_{0,i}$, on rejette l'hypothèse nulle et on accepte l'hypothèse alternative, si $\widehat{S}_i(x_{i,\text{obs}})$ n'est pas dans la zone de rejet, on conserve l'hypothèse nulle.

Une alternative à cette procédure de test est de prendre en compte à quel point la valeur observée est proche ou loin de la valeur seuil. Cette information supplémentaire est importante puisque si la valeur observée est loin du seuil, alors soit elle est loin de la zone de rejet et $H_{1,i}$ n'est pas assez vraisemblable entraînant la conservation de $H_{0,i}$; soit elle est au coeur de la zone de rejet et le test rejette $H_{0,i}$ avec grande confiance. Au contraire, si la valeur observée est proche du seuil, alors, quelle que soit la décision, elle est prise avec faible confiance (si le jeu de données est légèrement modifié, la nouvelle valeur observée peut faire basculer la décision du test). Pour cela, les statistiques de tests utilisées sont les *p-valeurs*. Si pour chaque x, $T_i(x) = \mathbb{P}_{P_i}\Big(X \in \widehat{R}_i\Big(\widehat{S}_i(x)\Big)\Big)$ pour $X \sim P_i$, alors les p-valeurs sont les variables aléatoires définies par :

$$\forall i \in \{1, \cdots, p\}, \quad \widehat{p}_i = \sup_{P_i \in \Theta_{0,i}} T_i(\tilde{S}_i), \quad \text{pour } \tilde{S}_i \text{ indépendant et de même loi que } \widehat{S}_i.$$

En d'autres termes, pour chaque $i \in \{1, \cdots, p\}$, $\widehat{p}_i(x_{i,\text{obs}})$ est la probabilité d'obtenir une observation au moins aussi extrême que $\widehat{S}_i(x_{i,\text{obs}})$ sous $H_{0,i}$. En remarquant que, pour un test de niveau α, le test rejette $H_{0,i}$ si et seulement si $\widehat{p}_i(x_{i,\text{obs}}) \leq \alpha$, on en déduit qu'au plus la p-valeur est faible, au plus $\widehat{S}_i(x_{i,\text{obs}})$ est au coeur de la zone de rejet et donc au plus la découverte $H_{1,i}$ est vraisemblable.

Utiliser les p-valeurs donne une procédure de test équivalente à la précédente. Cependant, leur écriture sous forme de probabilité permet de les utiliser formellement pour toutes lois $(P_1, \cdots, P_p)$ et pour n'importe quelle forme d'hypothèses de test. De plus, dans le cas où les P_i sont des lois continues, les p-valeurs ont une distribution uniforme sous l'hypothèse nulle. Cette propriété est massivement utilisée dans la littérature. Nous proposons la section 1.2.4 de [Durand, 2018] pour une revue des procédures de test lorsque les P_i ne sont pas continues.

- Tests multiples :
Nous disposons d'un jeu de p hypothèses nulles $(H_{0,1}, \cdots, H_{0,p})$ à tester contre p hypothèses alternatives $(H_{1,1}, \cdots, H_{1,p})$. Si les tests sont réalisés individuellement les uns des autres ($H_{0,i}$ est rejetée si et seulement si $\widehat{p}_i(x_{i,\text{obs}}) \leq \alpha$), alors tous contrôlent l'erreur de première espèce par α. Cependant, lorsque l'on considère les p décisions dans un ensemble global, la probabilité de rejeter une des hypothèses $H_{0,i}$ alors que toutes les hypothèses $H_{0,i}$ sont vraies n'est majorée que de $p\alpha$ dans le cas général, et de $1-(1-\alpha)^p$ si les $(X_1, \cdots, X_p)$ sont indépendantes les unes des autres (l'erreur totale est multipliée). Lorsque p est grand, ces quantités sont proches de 1 ou dépassent 1 : de faux rejets sont presque sûrement faits sur l'ensemble des tests. De plus, le contexte de grande dimension s'applique et en général, le nombre p_0 d'hypothèses nulles $H_{0,i}$ vraies est grand : peu de rejets sont à faire par rapport au nombre de tests p (hypothèse de parcimonie). Or, considérées séparément, les procédures de tests donnent un nombre moyen de faux rejets totaux égal à $p_0\alpha$, ce qui est donc trop grand. Ainsi, en grande dimension, les rejets sont basés sur des procédures contrôlant des fonctions de coût prenant en compte l'ensemble des tests simultanément : c'est le principe de correction des tests multiples.

De nombreuses applications des tests multiples ont vu le jour, notamment suite à l'explosion des données disponibles. Par exemple, en génomique, la technique des *puces à ADN* [Lenoir and Giannella, 2006] (voir sous-section 6.7), par laquelle les données réelles utilisées dans cette thèse ont été extraites, ont permis d'acquérir les niveaux d'expression de gènes à un instant donné pour tous les gènes d'un organisme simultanément. Les tests multiples peuvent être utilisés pour tester simultanément si les gènes se sont exprimés ou non à l'instant de l'extraction des données, pour comparer le niveau d'expression de tous les gènes simultanément entre deux phénotypes différents, ou pour rechercher la co-expression entre les gènes dans un phénotype donné par une étude paire par paire simultanée [Dudoit and van der Laan, 2008]. Les tests multiples sont utilisés dans d'autres domaines comme par exemple l'imagerie médicale, l'astrophysique ou l'industrie.

- Deux fonctions de coût : le FWER et le FDR :
On définit successivement :

- $\mathrm{P} = \Big\{ i \in \{1, \cdots, p\},\ H_{0,i} \text{ est rejetée} \Big\}$
- $\mathrm{N} = \Big\{ i \in \{1, \cdots, p\},\ H_{0,i} \text{ est conservée} \Big\}$
- $\mathrm{FP} = \Big\{ i \in \{1, \cdots, p\},\ H_{0,i} \text{ est rejetée à tort} \Big\}$
- $\mathrm{TP} = \Big\{ i \in \{1, \cdots, p\},\ H_{0,i} \text{ est rejetée à raison} \Big\}$
- $\mathrm{FN} = \Big\{ i \in \{1, \cdots, p\},\ H_{0,i} \text{ est conservée à tort} \Big\}$.

L'ensemble FP contient les faux rejets. Idéalement, on veut minimiser la taille de FP, tout en maximisant celle de TP. Un tel compromis est un réel défi. Le premier critère mis en place est le *Family-Wise Error Rate (FWER)* qui est la probabilité que la procédure produise au moins un faux rejet :

$$\mathrm{FWER} = \mathbb{P}\Big(\#FP > 0 \Big).$$

La *correction de Bonferroni* [Bonferroni, 1936, Dunn, 1961] consiste à prendre comme zone de rejet :

$$\widehat{R}_{\mathrm{Bonf}} = \left\{ i \in \{1, \cdots, p\}, \quad \widehat{p}_i \leq \frac{\alpha}{p} \right\}.$$

Dans ce cas, $\mathrm{FWER} \leq \alpha$, et ce pour n'importe dépendance sur la famille de p-valeurs $\Big(\widehat{p}_1, \cdots, \widehat{p}_p\Big)$. Ainsi, la probabilité de ne faire aucun faux rejet sur l'ensemble des tests est grande : $1 - \alpha$. Cependant, cette procédure est équivalente à réaliser chaque test au niveau $\frac{\alpha}{p}$. Comme $\frac{\alpha}{p} \ll \alpha$, la procédure est beaucoup plus conservative que celle consistant à traiter les tests individuellement les uns des autres. Ainsi, rejeter $H_{0,i}$ est plus rare et la discrimination entre les $H_{0,i}$ et les $H_{1,i}$ perd en capacité. Dans un contexte de grande dimension, le nombre d'hypothèses alternatives $H_{1,i}$ vraies est faible par rapport à p, ainsi, empêcher la détection de faux rejets empêche aussi fortement la détection de vrais rejets. Une solution est de s'autoriser une proportion de faux rejets parmi les rejets faits.

En ce sens, [Benjamini and Hochberg, 1995] propose de contrôler le proportion du taux de faux rejets parmi les rejets ou son espérance. Ainsi, sont définis formellement le *False Discovery Proportion (FDP)* et le False Discovery Rate (FDR) :

$$\mathrm{FDP} = \frac{\#FP}{\#P \vee 1},$$

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#FP}{\#P \vee 1} \right].$$

On a toujours $\mathrm{FDR} \leq \mathrm{FWER}$, avec égalité si $p_0 = p$: contrôler le FDR est moins conservatif que contrôler le FWER. Pour contrôler le FDR, une idée très généralement développée dans la littérature consiste à étudier les p-valeurs. En effet, pour garantir simultanément un ensemble

TP grand et un ensemble FP petit, les hypothèses nulles rejetées doivent correspondre aux plus petites des p-valeurs. Soit $\left(\widehat{p}_{(1)}, \cdots, \widehat{p}_{(p)}\right)$ la famille des p-valeurs rangées dans l'ordre croissant : $\widehat{p}_{(1)} \leq \cdots \leq \widehat{p}_{(p)}$. Il ne reste qu'à trouver un seuil $\hat{k}$ en dessous duquel les p-valeurs sont associées aux vraies hypothèses alternatives et au delà duquel les p-valeurs sont associées aux vraies hypothèses nulles. Pour cela, [Benjamini and Hochberg, 1995] propose la zone de rejet suivante : pour $\gamma = \{1, \cdots, p\} \mapsto \mathbb{R}^+$ fonction croissante quelconque,

$$\widehat{R}_{\text{B.H}} = \left\{ i \in \{1, \cdots, p\}, \quad \widehat{p}_{(i)} \leq \frac{\gamma(\hat{k})}{p}\alpha \right\} \tag{2.25}$$

où

$$\hat{k} = \max \left\{ k \in \{1, \cdots, p\}, \quad \widehat{p}_{(k)} \leq \frac{\gamma(k)}{p}\alpha \right\} \tag{2.26}$$

avec pour convention $R_{\text{B.H}} = \emptyset$ (c'est-à-dire $\#P = 0$) si $\left\{ k \in \{1, \cdots, p\}, \quad \widehat{p}_{(k)} \leq \frac{\gamma(k)}{p}\alpha \right\} = \emptyset$.

Théorème 2.3. *([Benjamini and Hochberg, 1995]).*
Soit $\gamma : \{1, \cdots, p\} \mapsto \mathbb{R}^+$ une fonction croissante quelconque et soit la zone de rejet $\widehat{R}_{B.H}$ (2.25). Supposons que les p-valeurs ont une distribution uniforme sur l'hypothèse nulle. Alors :

$$FDR \leq \alpha \frac{p_0}{p} \sum_{i=1}^{p} \frac{\gamma(i \wedge p)}{i(i+1)}.$$

Cette inégalité est optimale car elle est atteinte pour certaines distributions sur les p-valeurs et un α relativement petit [Guo and Rao, 2008]. Il ne reste plus qu'à trouver des fonctions γ tel que FDR $\leq \alpha$ (en moyenne, il y a au plus $\alpha\%$ de faux rejets parmi les rejets) tout en maximisant la taille de la zone de rejet. Dans cette direction, [Benjamini and Yekutieli, 2001] propose $\gamma(k) = \frac{k}{H_p}$ avec $H_p = 1 + \frac{1}{2} + \cdots + \frac{1}{p} \approx \log(p)$. Dans ce cas, pour n'importe quelle dépendance au sein de la famille de p-valeurs, FDR $\leq \alpha$. Ce choix de fonction donne de nouveau une procédure très conservative et il n'est pas possible d'élargir la zone de rejet en toute généralité.

En revanche, si la zone de rejet est élargie par $\gamma(k) = k$, alors FDR $\leq \frac{p_0}{p}\alpha \leq \alpha$ si les p-valeurs sont indépendantes [Benjamini and Hochberg, 1995], ou si les p-valeurs vérifient la propriété dite de *Weak Postive Regression Dependence Property (WPRDP)* [Benjamini and Yekutieli, 2001] où des dépendances très strictement contrôlées entre les p-valeurs sont tolérées.

- Quelques extensions :
Contrôler le FWER ou le FDR revient à satisfaire un contrôle de l'erreur de première espèce, mais aucune considération n'est faite sur l'erreur de seconde espèce. Plusieurs méthodes ont été proposées, soit en modifiant la fonction de coût, soit en considérant plusieurs fonctions de coût simultanément. Nous proposons une liste non-exhaustive dans ce paragraphe. Par exemple, [Lehmann and Romano, 2005,Romano et al., 2007] proposent de contrôler le *k-FWER*,

qui est, par relaxation du FWER, la probabilité que le nombre de faux rejets sur l'ensemble des tests soit supérieure ou égale à k. Dans une toute autre direction, [Benjamini and Hochberg, 2000] et [Storey, 2002] proposent d'estimer la valeur de p_0 pour obtenir une estimation de la fonction de coût FDR, et de calibrer ensuite la valeur de α. Pour cela, les p-valeurs sont supposées indépendantes les unes des autres. Sous cette dernière hypothèse, les p-valeurs sont des réalisations d'un échantillon ordonné de la loi uniforme. Par exemple, p_0 peut être estimé en $\widehat{p_0}$ par détection d'une rupture de comportement sur la famille $\widehat{p}_1 \leq \cdots \leq \widehat{p}_p$, et la région de rejet peut être alors calculée à l'aide de la nouvelle définition de $\hat{k}$

$$\hat{k} = \max\left\{k \in \{1, \cdots, p\}, \quad \widehat{p}_{(k)} \leq \frac{k}{\widehat{p_0}}\alpha\right\}. \tag{2.27}$$

Des résultats théoriques sont vérifiés. Par exemple, [Storey, 2002] prouve que le FDR associé est contrôlé asymptotiquement sous une condition de dépendance faible. Pour pallier l'hypothèse d'indépendance sur les p-valeurs, [Romano et al., 2008] propose un estimateur du FDR par bootstrap pour estimer le FDR. Celui-ci est alors contrôlé asymptotiquement sous une hypothèse d'interchangeabilité. Plutôt que le FDR, [Genovese et al., 2004] propose de contrôler le FDP, toujours en exploitant la propriété que $\Big(p_i | H_{0,i} \text{ est vraie}\Big) \sim \text{Unif}(0,1)$ et $\Big(p_i | H_{1,i} \text{ est vraie}\Big) \sim h$ pour h une distribution différente de $\text{Unif}(0,1)$. Ainsi, si F_f désigne la fonction de répartition de la loi f, alors la distribution marginale des p-valeurs est $G = \Pi_0 F_{\text{Unif}(0,1)} + (1 - \Pi_0) F_h$, qu'ils exploitent pour obtenir une estimation du FDP. Une alternative, proposée par [Storey et al., 2003] est le *pFDR* (pour *positive FDR*) défini par $\mathbb{E}\Big[\frac{\#FP}{\#P}\Big|\#P > 0\Big]$ pour remplacer le FDR, et les q-valeurs définies par $\text{q-value}_i := \inf_{t \leq t_{\alpha,i}} \text{pFDR}(t)$ où $t_{\alpha,i} = \min\Big(t \in [0,1],\ \mathbb{P}\Big(p_i \leq t \Big| H_{0,i} \text{ est vraie}\Big) = \alpha\Big)$ pour remplacer les p-valeurs. Comme les intervalles de confiance sont plus informatifs qu'une hypothèse de test rejetée ou non, [Benjamini and Yekutieli, 2005] propose de les contrôler à la place du FDR. Une toute autre méthode consiste à tenir compte que les tests peuvent ne pas avoir la même importance dans la procédure de tests multiples. Par exemple, [Holm, 1979] utilise les quantités $\frac{p_i}{w_i}$ au lieu des p_i où les w_i sont estimées au préalable par une procédure de statistique indépendante des tests multiples de sorte qu'un contrôle adaptatif (à p_0) du FWER soit obtenu ; [Genovese et al., 2006] utilise des poids aléatoires pour contrôler le FDR. La fonction de coût FDR est un critère global qui ne donne pas d'informations sur un test en particulier. C'est pourquoi, [Efron, 2005] propose de contrôler le *local FDR* plutôt que le FDR. Enfin, afin de contrôler l'erreur de première espèce et l'erreur de seconde espèce, [Genovese and Wasserman, 2002] propose de considérer le *False Negative Rate (FNR)* défini par :

$$\text{FNR} = \mathbb{E}\Big[FNP\Big]$$

où

$$\text{FNP} = \frac{\#FN}{\#N \vee 1}$$

Pour cela, ils minimisent fonction de coût FDP + λFNP où λ est un hyperparamètre estimé sur le jeu de données disponible afin que le FNR soit le plus petit possible sous le contrôle

FDR $\leq \alpha$.
Nous renvoyons le lecteur vers l'article [Roquain, 2011] où un résumé détaillé de ces approches est proposé.

Chapter 3

An overview of variable selection procedures using regularization paths in high-dimensional Gaussian linear regression

Abstract.

Current high-throughput technologies provide large amount of variables to describe a phenomenon and high-dimensional Gaussian linear regression is the one of the most-used statistical methods to identify active variables related to the response variable. In this article, we describe step-by-step the variable selection procedures built upon regularization paths, obtained by combining a regularization function, and an algorithm. These paths are either combined with a penalty to perform a model selection or with sampling strategies for variable identification. We perform a comparison study by considering three simulation settings. In all the settings, we evaluate the performance of construction of the regularization path (pROC-AUC), the prediction performance (MSE), and the relevance of the selected variables (recall, specificity, FDR). The results are the basis for recommendations on which method to use depending on the characteristics of the problem at hand. The regularization function Elastic-net provides most of the time better results than the ℓ_1 one and the lars algorithm has to be privileged as the GD one. ESCV leads to the best prediction performances. Bolasso and the knockoffs method are a judicious strategy to limit the selection of non active variables while ensuring selection of enough active variables. Conversely, the data-driven penalties considered in this review have to be avoided. As for Tigress and LinSelect, they are too conservative methods.

Keywords: High-dimensional, Gaussian linear regression, Variable selection, Regularization path, Neutral comparison study

This work was conducted in collaboration with Mélina Gallopin (Institute for Integrative Biologiy of the Cell (I2BC)).

Contents

3.1 Introduction

Recent scientific advances allow us to have access to large-scale data: the size of the data sets is exploding, as well as the complexity of each of them. For instance, in genomics, to describe the molecular activities, microarrays and RNA-sequencing technologies providing a quantification of the expression of all the genes simultaneously ease the study of their interactions. Genetic associations studies diversify by considering a large range of phenotypes including gene expression or proteomic and metabolomic data. In a statistical point of view, the number of parameters to estimate explodes and reduction of dimension is required to select only relevant variables and summarize the redundant information for a given model. In this review, we focus on the variable selection procedures in high-dimensional linear Gaussian regression model. The considered dataset with a number of variables p close to or slightly higher than the number of observations n is a real challenge since it hampers the use of the traditional estimation methods. A regularization of the cost function is required so that only a subset of variables is selected to explain the response variable.

In most reviews on variable selection in high-dimensional Gaussian linear regression, a focus is done on the construction of the regularization path. [Wainwright, 2009] provides a meticulous theoretical analysis of the ℓ_1 penalty function. In particular, for a given number of active variables, he discusses the choice of the number of observations to ensure asymptotic properties to recover these active variables. [Bühlmann and Mandozzi, 2014] compared several regularization functions in a simulation study by using semi-real datasets. In their simulation design they considered several numbers of observations, variables and active variables. They also modified the signal-to-noise ratio and considered two scenarios of variable correlations. The results of the different regularization functions are inspected with ROC curves and partial ROC curves when $0.5n$ and $0.9n$ variables are selected. [Wang et al., 2020] compared a large set of regularization functions with a simulation design similar to [Bühlmann and Mandozzi, 2014]. They evaluated both prediction and variable identification but the main difference with our investigations is that the model selection is only performed by a cross-validation procedure. Finally some reviews considered different contexts. [Wu and Ma, 2015] were interested in robust variable selection strategies when heavy-tailed errors and outliers in response variables exist. They discussed the different steps from the modification of the least squares function to the choice of the parameters for the model selection through a presentation about algorithms accounting for outliers. [Vinga, 2021] considered a variety of models from survival models to generalized linear models, frequently used in biomedical researches. [Desboulets, 2018] considered a wide range of model structures (linear, grouped, additive, partially linear and non-parametric) and discussed three main categories of algorithms for the variable selection.

Our review distinguishes itself from the previous ones since we propose an evaluation of both the regularization path construction and the choice of the final selected variables leading to 33 combinations. Moreover, we add in this review non-asymptotic methods of model selection which are generally not considered. To construct the regularization path, we test two regularization functions (Lasso [Tibshirani, 1996] and Elastic-Net [Zou and Hastie, 2005]) combined with two algorithms (LARS [Efron et al., 2004] and gradient descent algorithm [Friedman et al.,

2010]). Each regularization path provides a collection of variable subsets and to choose one of them, we compare model selection and variable identification approaches. On the one hand, the model selection uses penalization criteria of the least squares (eBIC [Chen and Chen, 2008], data-driven calibration strategies [Birgé and Massart, 2007, Lebarbier, 2005, Baudry et al., 2012, Arlot, 2019] and LinSelect [Baraud et al., 2009, Giraud et al., 2012]). On the other hand, the variable identification methods (ESCV [Lim and Yu, 2016], Bolasso [Bach, 2008], Stability Selection [Meinshausen and Bühlmann, 2010], Tigress [Haury et al., 2012] and the knockoffs method [Barber and Candès, 2015]) use sampling strategies to stabilize the selected variable subset while limiting the number of non active variables. Methods based on multiple testing procedures [Kos and Bogdan, 2020] and Bayesian approaches are not included in this review. We refer the readers to [Celeux et al., 2012] for an empirical comparison of frequentist and Bayesian points of view. Lastly, we assume no prior knowledge between interactions, spatial localization and chronological information and refer to [Bondell and Reich, 2012, Razaghi-Moghadam and Nikoloski, 2020] for such approaches.

After a description of the methods and their theoretical properties, we compare them in a simulation study by considering three settings. In the first one, the variables are independent and drawn from a Gaussian distribution. It allows a method comparison in the theoretical framework used to develop them. In the second setting, two structures of the correlation between variables are considered to evaluate how dependencies usually observed affect the methods. Observations are generated according to a Gaussian linear model, the most favorable case where assumptions broadly hold. Finally, the third setting mimics the biological complexity of transcription factor regulations. Observations are generated using the *FRANK* algorithm [Carré et al., 2017].

In these three settings, the method performances are evaluated for their prediction performance and for their ability to identification of the active variables. To discriminate active variables to the others, we use the pROC-AUC metric. We use the mean squared errors (MSE) to measure the prediction performance, and the recall, specificity and false discovery rate (FDR) metrics to assess the quality of the selected variables in terms of active variables. As [Bühlmann and Mandozzi, 2014, Wang et al., 2020], we notice that there is no unambiguous winner among all the studied approaches. Our idea is rather to offer recommendations for a judicious choice of a method according to the application. In particular, Elastic-Net should be preferred to Lasso for the regularization function, as well as the lars algorithm. Moreover, to ensure a prediction ability, ESCV and the knockoffs seem to be the most judicious choices. If the goal is to recover the active variables, ESCV and eBIC are preferable whereas Bolasso, the knockoffs and LinSelect should be privileged to limit the non active variables in the selected subset.

3.2 Methods

3.2.1 Statistical framework

We consider a Gaussian linear regression model where the response variable Y is explained by a linear combination of p variables $X = (X_1, \ldots, X_p)$:

$$Y = X\beta^* + \varepsilon$$

where $\beta^* \in \mathbb{R}^p$ is the unknown vector and ε follows a centered Gaussian distribution with an unknown variance denoted σ^2. To estimate the parameters β^* and σ^2, n independent observations are available: for $i \in \{1, ..., n\}$, we observe $y_i \in \mathbb{R}$ and the variables $(x_{i1}, ..., x_{ip}) \in \mathbb{R}^p$.
For the sequel, $||.||_2$ designs the usual Euclidean norm on $\mathbb{R}^n$. The norms $|.|_0$, $|.|_1$ and $||.||$ are respectively the usual norms 0, 1 and 2 on $\mathbb{R}^p$.
We consider high-dimensional regression where $p \sim n$ or $p \gg n$, preventing the traditional least squares estimation. So an additive regularization corresponding to a sparsity assumption is imposed: only a small number of variables explains the response variable. For $t > 0$, the criterion becomes:

$$\min_{\beta \in \mathbb{R}^p : |\beta|_0 \leq t} ||Y - X\beta||_2^2$$

where $|\beta|_0$ is the number of non-zero coefficients of β and $||.||$ denotes the standard Euclidean norm on $\mathbb{R}^n$. Its associated Lagrangian form is for $\lambda > 0$:

$$\min_{\beta \in \mathbb{R}^p} \left\{ ||Y - X\beta||_2^2 + \lambda |\beta|_0 \right\} \tag{3.1}$$

The proof of the equivalence and the link between t and λ is provided in [Tibshirani, 1996]. On the one hand, a large value of λ gives a small subset of variables (assumption of sparsity satisfied) but corresponding to a linear adjustment far from the response variable. On the other hand, a small value of λ gives a large subset of variables (assumption of sparsity not satisfied) but corresponding to a linear adjustment close to the response variable. Determining the hyperparameter λ is one of the major issues and the challenge lies in its calibration to adjust a trade-off between sparsity and a good linear adjustment. Unfortunately, the criterion is non-convex, so the existence and the uniqueness of the solution are not guaranteed. Hence, as presented in [Vinga, 2021], (3.1) is generalized as:

$$\min_{\beta \in \mathbb{R}^p} \left\{ ||Y - X\beta||_2^2 + \lambda F(\beta) \right\} \tag{3.2}$$

where F is a continuous and convex regularization function which guarantees the existence of a minimum for any λ.

3.2.2 Regularization functions

The Lasso procedure [Tibshirani, 1996] considers the ℓ_1-norm:

$$F(\beta) = |\beta|_1 = \sum_{j=1}^{p} |\beta_j|$$

Lasso is a good approximation of (3.1. If λ is well chosen, it provides a consistent estimator of β. This procedure achieves the best trade-off between regularity (convexity, reasonable computationally solving) and sparsity for independent variables. However, when some variables are correlated, Lasso tends to select randomly only one of them rather than selecting none or all of them. A solution is the Adaptative Lasso procedure [Zou, 2006] where each variable is properly weighted. There is also the Ridge procedure [Hoerl and Kennard, 1988]:

$$F(\beta) = ||\beta||^2 = \sum_{j=1}^{p} \beta_j^2$$

In addition to taking variable dependencies into account, Ridge provides a strictly convex and derivable optimization problem with an explicit estimator of β^*. However, this estimator is not sparse. Another alternative is the Elastic-Net procedure [Zou and Hastie, 2005] which balances the sparsity control with variable dependencies inclusion through the parameter α:

$$F(\beta) = (1-\alpha)|\beta|_1 + \alpha||\beta||^2$$

When prior knowledge on variable dependencies is available, there exist other regularization functions, not considered here: the Group Lasso [Yuan and Lin, 2006], Overlap Group Lasso [Jacob et al., 2009], Hierarchical Group Lasso [Zhao et al., 2009] and fused Lasso [Tibshirani et al., 2005].

3.2.3 Regularization path construction for Lasso and Elastic-Net

The optimization problem (3.2) has generally no explicit solution and must be solved by a computational approach by considering a grid Λ of λ. A first algorithm is LARS [Efron et al., 2004]. Briefly speaking, the first subset contains only one variable: the variable X_j which has the largest absolute correlation with Y. The second model contains exactly two variables: X_j and the variable X_k with $k \neq j$ being the most correlated variable with the residuals of the regression of Y on X_j. One variable is added at each step and each step corresponds to a value of λ. A second algorithm [Friedman et al., 2010] is based on the gradient descent method. This algorithm constructs a regular grid Λ of a given size by starting with the largest λ corresponding to the first nonempty variable subset. Then, a variable subset is obtained for each λ of this grid by solving (3.2) with the cyclical coordinate descent method.

LARS gives an exact solution of the optimization problem whereas the gradient descent method provides a proxy. LARS leads to nested subsets, an important property for theoretical considerations, whereas the gradient descent method gives independent solutions along the grid.

Consequently, LARS gives at most one subset per dimension whereas the gradient descent method may provide a richer collection.

3.2.4 Model selection

Whatever the chosen algorithm, a collection of variable subsets $(m_\lambda)_{\lambda\in\Lambda}$ is obtained. Each m_λ is associated with an estimator of β^*, however it is known to be biased [Meinshausen, 2007] and is usually replaced with the least-squares estimator calculated only on the variables of m_λ [Connault, 2011]. It is denoted $\hat{\beta}_\lambda$ and its dimension is denoted D_λ, corresponding to the number of variables in m_λ. To select the best subset $m_{\hat{\lambda}}$, model selection approaches consist of minimizing a penalized loss function in $\lambda \in \Lambda$:

$$\gamma(m_\lambda) + \text{pen}(n, p, D_\lambda) \tag{3.3}$$

The loss function, $\gamma(m_\lambda)$, quantifying the quality of the model fit is either the least-squares function $||Y - X\hat{\beta}_\lambda||_2^2$ or the deviance $-2\log(L(Y, X; \hat{\beta}_\lambda, \hat{\sigma}_\lambda^2))$, where L is the likelihood function calculated with $\hat{\beta}_\lambda$ and $\hat{\sigma}_\lambda^2$, the empirical estimators associated to the model m_λ. The penalty function $\text{pen}(n, p, D_\lambda)$ accounts for the model complexity and the characteristics of the sample: higher the penalty values, smaller the number of selected variables and farther the linear combination $X\hat{\beta}_\lambda$ to the response variable Y.

Asymptotic criterion The first criteria are asymptotic: theirs properties are verified when the sample size n tends to infinity. In this review, we focus on the more recent asymptotic criterion, called eBIC [Chen and Chen, 2008], used to get a consistent estimator by penalizing the deviance by:

$$\text{pen}_{\text{eBIC}}(n, p, D_\lambda) = D_\lambda \log(n) + 2\delta \log(\binom{p}{D_\lambda}) \tag{3.4}$$

where δ is a value in $[0, 1]$.

Non-asymptotic criteria In a practical consideration, a fixed real or simulated dataset is available. In this case, having estimator properties for n going to infinity has no sense (see [Giraud et al., 2012]) and applying criteria with properties confirmed for any fixed sample n size is more relevant. Introduced by Birgé and Massart [Birgé and Massart, 2001], the goal of non-asymptotic criteria is to achieve the risk oracle:

$$\inf_{\lambda\in\Lambda} \mathbb{E}[\ ||X\beta^* - X\hat{\beta}_\lambda||_2^2]$$

and instead of getting asymptotic equality of the kind

$$\mathbb{P}\left(\lim_{n\to+\infty} \frac{\mathbb{E}[\ ||X\beta^* - X\hat{\beta}_{\hat{\lambda}}||_2^2]}{\inf_{\lambda\in\Lambda} \mathbb{E}[\ ||X\beta^* - X\hat{\beta}_\lambda||_2^2]} = 1\right) = 1$$

they get an inequality holding for any value of n:

$$\mathbb{E}[\ ||X\beta^* - X\hat{\beta}_{\hat{\lambda}}||_2^2] \leq C_n \inf_{\lambda\in\Lambda} \{\mathbb{E}[\ ||X\beta^* - X\hat{\beta}_\lambda||_2^2]\ \} + R_n, \tag{3.5}$$

where $C_n \approx 1$ at least for n large and R_n is small comparable to the risk oracle $\inf_{\lambda \in \Lambda}\{\mathbb{E}[\ ||X\beta^* - X\hat{\beta}_\lambda||_2^2]\ \}$. The selection procedure which satisfies (3.5) is the same as for an asymptotic criterion: the selected model is the minimizer of Equation (3.3) where the loss function is the least-squares function and two penalty functions are available. The first penalty is a data-driven calibration function [Birgé and Massart, 2007]

$$\text{pen}_{\text{Data-driven}}(n, p, D_\lambda) = 2\kappa D_\lambda \Big(2.5 + \log\big(\frac{p}{D_\lambda}\big)\Big) \tag{3.6}$$

where the constant 2.5 has been calculated in a context of detection of changepoints in a signal [Lebarbier, 2005]. The constant κ is calibrated from the sample. Two strategies are proposed: the first one is the slope heuristics, assuming that the least-squares function is linear in $D_\lambda \left(2.5 + \log(\frac{p}{D_\lambda})\right)$ as soon as D_λ is large enough (see Figure 2 of [Baudry et al., 2012]). The constant κ equals the estimated slope. The second strategy is the dimension jump, assuming the existence of κ^* such that for all the values smaller than κ^*, the associated model has a very high dimension, whereas for all the values greater than κ^*, the associated model has a reasonable dimension (see Figure 1 of [Baudry et al., 2012]). The constant κ equals κ^*. For more practical and theoretical details, we refer the reader to [Baudry et al., 2012] and the survey [Arlot, 2019]. The second penalty function is LinSelect proposed in [Baraud et al., 2009] and generalized for a high dimensional context in [Giraud et al., 2012]. It is built from the empirical estimator of the variance onto each m_λ:

$$\text{pen}_{\text{LinSelect}}(n, p, D_\lambda) = 1.1 \times \frac{n - D_\lambda}{n - D_\lambda - 1} \Psi\left(D_\lambda + 1, n - D_\lambda - 1, e^{-L_\lambda}\right) \tag{3.7}$$

where the L_λ are weights satisfying some properties and the function $\Psi[D, N, q]$ is the unique solution of the equation:

$$\phi\left[D, N, \Psi(D, N, q)\right] = q$$

where $\phi[D, N, x]$ is defined for $x \geq 0$:

$$\phi[D, N, x] = \frac{1}{D}\mathbb{E}\left[\max\left(0, \chi_D^2 - x\frac{\chi_N^2}{N}\right)\right]$$

for χ_D^2 and χ_N^2 two independent χ^2 random variables with degrees of freedom D and N respectively.

Both penalties are based on the theory developed in [Birgé and Massart, 2001] and define an explicit penalty function when the variance is unknown. LinSelect deals with the empirical variance leading to the constant 1.1 in Equation (3.7), whereas data-driven penalties prefer to calibrate the variance from the sample. This last approach is an heuristic tested and verified in various frameworks without theoretical properties [Baudry et al., 2012].

In our simulation study, 16 methods are defined by the combination of a regularization function (Lasso or Elastic-net) with an algorithm (LARS or the gradient descent method) and a penalty function (eBIC, LinSelect or 2 data-driven penalties).

3.2.5 Variable identification

The high-dimensional framework usually leads to an unstable result: addition, suppression, or modification of some observations could radically change the subset of selected variables. For prediction, different sets of variables can give the same prediction performance. However, when the objective is the identification of the active variables, this instability is a drawback.

To circumvent this sampling uncertainty, the idea is to replicate the procedure on perturbed data generated from the original dataset. Cross-validation [Allen, 1974, Stone, 1974] consists in splitting K times the original dataset into a training and a test set. For each $k = \{1, ..., K\}$ and each $\lambda \in \Lambda$, the training set is used to calculate $\hat{\beta}^k_\lambda$ and the test set is used to evaluate the mean squared error. The selected model minimizes in λ the mean squared error. Applying cross-validation in high dimension is computationally expensive and known to be unstable, ESCV [Lim and Yu, 2016] estimates the instability along the regularization path instead:

$$\frac{\frac{1}{K}\sum_{k=1}^{K}||X\hat{\beta}^k_\lambda - \frac{1}{K}\sum_{\ell=1}^{K}X\hat{\beta}^\ell_\lambda||_2^2}{||\frac{1}{K}\sum_{\ell=1}^{K}X\hat{\beta}^\ell_\lambda||_2^2}$$

Sampling strategy is also a solution. Two widely used approaches are Bolasso and Stability Selection proposed by [Bach, 2008, Meinshausen and Bühlmann, 2010], which mainly differ in their sampling strategy. Bolasso generates samples of n data uniformly chosen with replacement among the n original observations, whereas the Stability Selection strategy generates samples of $\lfloor \frac{n}{2} \rfloor$ distinct data randomly chosen. Moreover, in the Stability Selection method, to limit the subsampling effects, when a sample is generated, its complement is also taken into account and a random perturbation can be added in the regularization function $F(\beta)$:

$$\sum_{j=1}^{p}\frac{|\beta_j|}{W_j}$$

where $W_j \sim \mathcal{U}([\alpha, 1])$ with $\alpha > 0$. For these two approaches, we consider either a given grid or a grid varying with each sample. Sampling strategies allow one to get an occurrence frequency of each variable for each λ of the grid and those having the highest occurrence frequencies are retained to constitute the final variable subset. Tigress method [Haury et al., 2012] modifies the calculation of the occurrence frequency by averaging also on the LARS steps.

The last type of variable identification method is the knockoffs method [Barber and Candès, 2015]. It controls the FDR. This method starts with the construction of a matrix $\tilde{X}$ such that $\tilde{X}$ and X have the same covariance structure with $\tilde{X}_j$ the less correlated to X_j. The matrix $\tilde{X}$ is designed through linear algebra tools (see [Barber and Candès, 2015, Candès et al., 2016] for details). Then, a regularization path is constructed on the augmented matrix $X\tilde{X}$ of size $n \times 2p$ where the active variables are expected to be selected very earlier than their copy. Let denote

$$W_j = \max\left(Z_j, \tilde{Z}_j\right) \times \text{sign}\left(Z_j - \tilde{Z}_j\right) \tag{3.8}$$

where Z_j and $\tilde{Z}_j$ correspond to the largest λ for which X_j and $\tilde{X}_j$ respectively are selected. A positive value of W_j states that X_j is selected before its copy $\tilde{X}_j$ and a large positive value indicates that X_j is selected rapidly. Let q be the target FDR, the final variable subset is composed by the X_j such that $W_j \geq T$ with:

$$T = \min \left\{ t \in \mathcal{W}, \frac{1 + \#\{j : W_j \leq -t\}}{\min\left(1, \#\{j : W_j \geq t\}\right)} \leq q \right\}$$

where $\mathcal{W} = \{|W_j|, j = 1, ..., p\} \setminus \{0\}$. If the set is empty, T is set to infinity.

In our simulation study, 17 methods are defined for variable identification. When the same samples are used for all the λ of the grid, 8 methods are defined by the combination of the sampling strategy (Bolasso or Stability Selection) with a regularization function (Lasso or Elastic-net) and an algorithm (LARS or the gradient descent method). When the grid Λ is fixed, the sampling strategy is performed for each λ of the grid, which implies using the gradient descent algorithm, hence 4 methods are defined. We don't investigate the impact of the presence of a random perturbation in the regularization function $F(\beta)$: we set $\alpha = 1$. Furthermore, we compare Tigress, the knockoffs method and ESCV. For the last two, a gradient descent algorithm is used with either Lasso or Elastic-net.

3.3 Comparison study

3.3.1 Three simulation settings

The design of the simulation study is directly linked to the high dimensional regression model and is composed of three simulation settings. The study of the behavior of statistical methods on different dependency structures on variables is at the heart of this work. On the one hand, variables are independent with no connection (under an independent design). On the other hand, variables have some dependency structures (under a Gaussian graphical model or under a dynamic process).

Simulation under independent design This is the simplest setting where the high-dimensional framework is the single handicap [Fan et al., 2014, Wang et al., 2020]. The matrix X is simulated by concatenation of p independent standard Gaussian vectors of size n. The number of non-zero coefficients of the vector β^* is generated from a uniform variable on integers between 10 and 15. Theirs values are generated from a uniform distribution between 0.5 and 2 and the response variable Y satisfies $Y = X\beta^* + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, I_n)$.

Simulation under a Gaussian graphical model In [Meinshausen et al., 2006], an equivalence is obtained between the network inference by Gaussian Graphical model and supports recovering by Gaussian linear regressions model. The node i models the variable X_i and a link between the node i and the node j models a dependency link between variable X_i and variable X_j meaning that either X_j is on the true support of the regression when X_i is the regressed

variable; or X_i is on the true support of the regression when X_j is the regressed variable. Moreover, data generated from a network structure give dependent variables and this is the most favorable situation where assumptions on our model broadly hold. In this direction, our data with dependency structure are simulated under the Gaussian Graphical model and recovering links between variables is equivalent to recover supports by solving one Gaussian linear regression per variable. In this case, this last variable is the regressed variable. More precisely, a sample of size n is generated from a $(p+1)$ multivariate centered Gaussian distribution with covariance matrix Σ, where the dependency structure is encoded in the precision matrix Σ^{-1}: if its coefficient $(j, j^{'})$ is non-zero, it means that the variables X_j and $X_{j'}$ interact when all other variables are given. We refer to [Lauritzen, 1996] and [Córdoba et al., 2019] for more information. The response variable Y is chosen as a column of the $(p+1)$ multivariate centered Gaussian, whereas in the previous papers [Zou and Yuan, 2008, Wang et al., 2012, Peng and Wang, 2015, Wang et al., 2020], the response variable is usually simulated once the matrix X is fixed. This choice has been motivated by applications where the status of the response and active variables are commonly similar. Then, Y is removed from the matrix X so that X is a matrix of size $n \times p$. We consider two types of dependency:

- *Cluster* design: the precision matrix is simulated as a block diagonal matrix with B blocks of equal size where B divides $(p+1)$. The response variable Y is defined as the first variable.
- *Scale-free* design: a few variables have a lot of neighbors while all the others have a few ones. We consider two response variables corresponding to the variables having the highest and the smallest number of neighbors. These simulation designs are named *scale-free-max* and *scale-free-min* respectively.

Simulation under a dynamical process It is the most realistic setting, based on the gene regulatory networks-simulating engine called *FRANK* [Carré et al., 2017]. This algorithm simulates large gene regulatory networks with characteristics observed *in vivo*: variables are categorized into a set of transcription factors that activates or inhibits a set of target genes. We use FRANK with only transcription factors in order to compare the results with those from the Gaussian settings. We consider two possible response variables corresponding to the variables having the highest and the smallest number of neighbors. These simulation designs are named *FRANK-max* and *FRANK-min* respectively.

Simulation parameters For the 6 scenarios: independent, cluster, scale-free-max, scale-free-min, FRANK-max and FRANK-min, we set $n = 150$ and $p = 199$; 100 samples of size $2n$ are generated to create a training set of size n used for the estimation and a validation set of size n used to evaluate the method. For each sample, the variables are primarily centered and scaled. For the simulation under the Gaussian graphical model, we use the function *huge.generator* from the R package *huge* (version 1.3.4.1). For the cluster design, the block number B equals 5 and the probability of connection within a component is set to the default value 0.3. For simulation under a dynamical process, we use the online FRANK algorithm available on the

website `https://m2sb.org/?page=FRANK` with p transcription factors and $2n$ observations. The number of eigenvalues of the matrix on the unit circle is fixed to 2 and the minimum and maximum of sparsity are set to 1 and 50. Other parameters are set to default values.

For the LARS algorithm, we use the function *enet* of the R package *elasticnet* (version 1.1.1). The maximal number of steps to define the grid size is the default value $50 \times \min(p, n-1)$. For the gradient descent method, we use the function *glmnet* of the R package *glmnet* (version 3.0) and set the grid size at 1000. Both functions propose the Lasso and elastic-net regularization functions. We set $\alpha = 0.5$ for Elastic-Net regularization.

To perform model selection, we implement eBIC setting δ to 1. For the non-asymptotic criteria, LinSelect is implemented in the function *tuneLasso* of the R package *LINselect* (version 1.1.3). The data-driven penalties are calculated by using the function *capushe* of the R package *capushe* (version 1.1.1). The parameters are set to the default values except for the minimum percentage of points for the plateau selection set to 0.1.

To perform variable identification, we use the function *escv.glmnet* of the R package *HDCI* (version 1.0.2) for the ESCV strategy, with a number of groups $K = 10$. We implement Bolasso and Stability Selection with 100 samples. In this work, we do not investigate the impact of the presence of a random perturbation in the regularization function $F(\beta)$, so we fix $\alpha = 1$. A variable is selected when its occurrence frequency is higher than 0.8. To apply Tigress, we use the function *tigress* of the R package *tigress* (version 0.1.0) with 50 steps for the LARS algorithm. For the knockoffs method, we use the function *knockoff.filter* with option *create.second_ order* of the R package *knockoff* (version 0.3.2), we calculate the W_j with the function *stat.lasso_ lambdasmax* and set the FDR to 0.1.

3.3.2 Evaluation metrics

All the metrics presented below are calculated on each test set and discussed averaged on 100 test sets.

First, we evaluate the performance of regularization path constructions. We use the partial area under the receiver operating characteristic curve (pROC-AUC) where the x-axis is the proportion of selected non active variables among the non active variables and the y-axis is the proportion of selected active variables among the active variables. A high value of pROC-AUC states that the regularization path is able to discriminate the active variables from the others. The lengths of the regularization path differ between the regularization and algorithms uses. To fairly compare them, the pROC-AUC are calculated by truncating the values of the x-axis at the largest value common to all the regularization paths.

Second, we evaluate the prediction performance of the methods by the mean squared errors (MSE) calculated on each set test $(\tilde{Y}, \tilde{X})$:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Y}_i - (\tilde{X}\hat{\beta}_{\hat{\lambda}})_i\right)^2 \tag{3.9}$$

where $\hat{\beta}_{\hat{\lambda}}$ is the estimator of β^* selected from the associated training set. As data are centered and scaled, a MSE value lower than 1 means that the method has a prediction ability: the selected variables predict Y better than an empty set of variables.

Finally, we evaluate the variable identification by using three metrics: the recall (the proportion of the selected active variables among the active variables), the specificity (the proportion of the non active variables not selected among the non active variables) and the false discovery proportion (the proportion of selected non active variables among the selected variables). The recall and the specificity respectively control the number of active variables which are selected and the number of non active variables which are non selected, while the false discovery proportion metric evaluates the quality of the selected variables subset. By averaging on the test sets, the false discovery proportion becomes the FDR. Since the objective is to limit the selection of non active variables while selecting as many active variables as possible, recall and specificity are expected to be close to 1 while the FDR is expected to be low or close to the fixed level for the knockoffs method.

3.4 Results

This section presents the analyses of the results by applying all possible combinations of regularization path constructions and variable selection methods. Subsections 3.4.1- 3.4.7 are devoted to the 4 scenarios: independent, cluster, scale-free-max, scale-free-min. Subsection 3.4.8 is devoted to the 2 scenarios FRANK-max and FRANK-min. Boxplots of results in Subsections 3.4.1- 3.4.6 are provided in Appendix 3.7 of the Chapter 3. Boxplots of results in Subsections 3.4.7 and 3.4.8 are provided in supplementary material available in [1].

For the sequel, the notation *ind* is the diminutive of the independent design. The *GD*, *E-Net* and ℓ_1 denote respectively the gradient descent algorithm, the Elastic-Net regularization function and the Lasso regularization function. The slope heuristics method and the dimension jump are named respectively by *slope* and *jump*. Bolasso and Stability Selection are named respectively by *bol* and *ss*. Lastly, *grid* and *sub* denote the strategy, respectively when grids are first generated and the strategy when samples are first generated.

In the sequel, we discuss the medians of the results obtained by each method per simulation design. For each Subsection from 3.4.2 to 3.4.6, the first paragraph is a brief presentation of the metric; the next two concern the results obtained respectively from the independent design and the other three; the last one stands for a conclusion.

3.4.1 Size of the variable subsets

Table 3.1 summarizes the active variable number per simulation setting. The sizes of the estimated supports of β^* do not take extreme values, the sparsity hypothesis is respected. The

[1] `https://sites.google.com/view/placroix/research`

Active variable number	ind	cluster	scale_free_max	scale_free_min
mean *(sd)*	12.59 *(1.76)*	11.63 *(2.75)*	31.41 *(9.70)*	1 *(0)*

Table 3.1: Active variable number

independent design is a benchmark in this comparison study. Indeed, the statistical properties of each method must be verified on the independent structure. Moreover, the active variable numbers mean of the independent design is close to that of cluster one (around 12, relatively small). This choice allows evaluating, through the cluster design, the impact of the presence of a dependency structure between the variables. The two other designs test the effects of different dependency structures. Concerning the scale-free design, the support size always equals 1 for scale-free-min. In contrast, many support sizes are considered for scale-free-max: the support size is 31.41 on average with a high standard deviation (9.70), which allows to fully investigate the method performances under various support sizes.

Tables 3.2 and 3.3 summarize the number of selected variables per method and per simulation setting. Concerning the model selection procedures (Table 3.2), LinSelect is very conservative since the mean values are always between 1 and 2. eBIC seem to be also conservative since the means are around 14, 1, 7 and 1 for number of active variables around respectively 13, 12, 31 and 1. For these two penalties, the standard deviation values are small. Concerning the data-driven penalties, values are always significantly larger than the number of active variables; the dimension jump gets closer to the number of active variables compared to the slope heuristic. Moreover, the large standard deviation values show instability in the size of the selected variables subset for both penalties. Lastly, the lars algorithm and the E-Net regularization seem to be privileged. Concerning the variable identification methods (Table 3.3), almost all the standard deviation values are small. Tigress is a very conservative method with a size of selected variables subsets not exceeding 2.14 on average. The knockoffs method leads to very small subsets for both cluster and scale-free-min designs but selects around 12 and 26 variables for respectively the independent and the scale-free-max designs. Concerning the Bolasso and Stability Selection, mean values remain quite low (smaller than 13) and Stability Selection selects slightly more variables. The best combination to get closer to the number of active variables is the lars algorithm with the E-Net regularization and when samples are first generated. The method that achieves the closest values to the number of active variables is ESCV.

3.4.2 Discrimination of the active variables from the others

In this subsection, we evaluate the performance of regularization path constructions from either the ℓ_1 regularization, or the E-Net regularization and with either the lars algorithm or the GD one. To compare the ability to select the active variables first, we evaluate pROC-AUC of these

Selected variables number	ind	cluster	scale_free_max	scale_free_min
GD_E-Net_ebic (mean) *(sd)*	14.34 *(2.84)*	1.27 *(0.63)*	7.53 *(2.30)*	1.03 *(0.17)*
GD_E-Net_slope (mean) *(sd)*	36.80 *(16.64)*	62.69 *(34.98)*	60.63 *(21.47)*	57.99 *(36.28)*
GD_E-Net_jump (mean) *(sd)*	25.83 *(12.23)*	23.37 *(17.13)*	42.20 *(20.19)*	24.38 *(19.70)*
GD_E-Net_linselect (mean) *(sd)*	1.07 *(0.33)*	1.32 *(1.12)*	1.59 *(1.51)*	1.04 *(0.20)*
GD_ℓ_1_ebic (mean) *(sd)*	13.50 *(2.21)*	1.27 *(0.63)*	7.67 *(2.40)*	1.03 *(0.17)*
GD_ℓ_1_slope (mean) *(sd)*	35.69 *(16.12)*	62.94 *(35.49)*	62.61 *(21.30)*	59.67 *(35.06)*
GD_ℓ_1_jump (mean) *(sd)*	28.23 *(11.93)*	23.85 *(17.61)*	42.49 *(20.16)*	24.12 *(18.67)*
GD_ℓ_1_linselect (mean) *(sd)*	1.07 *(0.33)*	1.30 *(1.02)*	1.53 *(1.49)*	1.04 *(0.20)*
lars_E-Net_ebic (mean) *(sd)*	14.80 *(3.76)*	1.25 *(0.61)*	7.43 *(2.43)*	1.02 *(0.14)*
lars_E-Net_slope (mean) *(sd)*	20.81 *(10.43)*	23.15 *(15.93)*	45.33 *(17.30)*	22.09 *(14.17)*
lars_E-Net_jump (mean) *(sd)*	23.47 *(13.39)*	23.34 *(13.89)*	43.78 *(15.88)*	22.86 *(13.69)*
lars_E-Net_linselect (mean) *(sd)*	1.04 *(0.28)*	1.23 *(0.76)*	1.55 *(1.51)*	1.03 *(0.17)*
lars_ℓ_1_ebic (mean) *(sd)*	13.50 *(2.21)*	1.26 *(0.61)*	7.71 *(2.36)*	1.02 *(0.14)*
lars_ℓ_1_slope (mean) *(sd)*	61.65 *(29.42)*	54.34 *(32.72)*	67.72 *(24.92)*	53.51 *(33.65)*
lars_ℓ_1_jump (mean) *(sd)*	32.30 *(15.33)*	25.13 *(19.54)*	44.72 *(22.14)*	23.84 *(18.35)*
lars_ℓ_1_linselect (mean) *(sd)*	1.04 *(0.28)*	1.26 *(0.81)*	1.46 *(1.44)*	1.03 *(0.17)*

Table 3.2: Selected variables number for model selection methods

four combinations of regularization and algorithm choices.

Figure 3.1 summarizes the pROC-AUC values of the different regularization paths. They are calculated by truncating the values of the x-axis at the largest value common to all the regularization paths. Hence, the reference value is not 1, which is unattainable because of the truncation.

We observe that the highest values for the independent framework are obtained with the lars

Selected variables number	ind	cluster	scale_free_max	scale_free_min
GD_bol_grid_E-Net (mean) *(sd)*	4.23 *(1.28)*	0.20 *(0.45)*	1.00 *(0.86)*	0.01 *(0.10)*
GD_bol_grid_ℓ_1 (mean) *(sd)*	4.08 *(1.32)*	0.10 *(0.36)*	0.46 *(0.59)*	0.00 *(0.00)*
GD_bol_sub_E-Net (mean) *(sd)*	2.48 *(1.02)*	0.17 *(0.43)*	0.42 *(0.57)*	0.01 *(0.10)*
GD_bol_sub_ℓ_1 (mean) *(sd)*	2.34 *(0.99)*	0.08 *(0.31)*	0.23 *(0.42)*	0.00 *(0.00)*
GD_escv_E-Net (mean) *(sd)*	11.61 *(4.42)*	6.81 *(8.71)*	22.42 *(11.18)*	4.64 *(10.37)*
GD_escv_ℓ_1 (mean) *(sd)*	11.59 *(4.41)*	6.86 *(8.62)*	22.42 *(11.13)*	5.07 *(10.46)*
GD_knockoffs_E-Net (mean) *(sd)*	12.11 *(4.68)*	0.00 *(0.00)*	25.74 *(8.44)*	0.00 *(0.00)*
GD_knockoffs_ℓ_1 (mean) *(sd)*	12.11 *(4.68)*	0.00 *(0.00)*	25.74 *(8.44)*	0.00 *(0.00)*
GD_ss_grid_E-Net (mean) *(sd)*	6.18 *(1.74)*	0.49 *(0.70)*	2.84 *(1.27)*	0.19 *(0.42)*
GD_ss_grid_ℓ_1 (mean) *(sd)*	6.16 *(1.70)*	0.33 *(0.59)*	1.73 *(1.10)*	0.09 *(0.29)*
GD_ss_sub_E-Net (mean) *(sd)*	3.89 *(1.23)*	0.34 *(0.59)*	1.64 *(1.05)*	0.15 *(0.39)*
GD_ss_sub_ℓ_1 (mean) *(sd)*	3.81 *(1.20)*	0.26 *(0.50)*	1.04 *(0.90)*	0.07 *(0.26)*
lars_bol_sub_E-Net (mean) *(sd)*	7.15 *(1.35)*	1.48 *(1.65)*	9.19 *(2.21)*	0.43 *(0.54)*
lars_bol_sub_ℓ_1 (mean) *(sd)*	0.00 *(0.00)*	0.00 *(0.00)*	0.04 *(0.20)*	0.00 *(0.00)*
lars_ss_sub_E-Net (mean) *(sd)*	9.11 *(1.35)*	2.79 *(1.97)*	12.64 *(2.25)*	0.99 *(0.80)*
lars_ss_sub_ℓ_1 (mean) *(sd)*	0.00 *(0.00)*	0.00 *(0.00)*	0.16 *(0.42)*	0.00 *(0.00)*
lars_tigress_ℓ_1 (mean) *(sd)*	2.14 *(0.89)*	0.31 *(0.56)*	0.40 *(0.59)*	0.14 *(0.35)*

Table 3.3: Selected variables number for variable identification methods

algorithm: almost 0.99 for the E-Net regularization and 0.73 for the ℓ_1 one. Concerning the GD algorithm, the results are similar for both regularization functions: around 0.52.

For correlation structure settings, a decrease in values is observed compared to the independent setting for the lars algorithm: the E-Net regularization achieves 0.84 and 0.98 for respectively

cluster and scale-free-max designs but 1 for scale-free-min design; the ℓ_1 regularization achieves 0.55 and 0.64 for respectively cluster and scale-free-max designs but 0.73 for scale-free-min design. The GD algorithm's values are similar for the cluster and scale-free-max designs but achieve 0.72 for the scale-free-min one. Conclusions are the same as independent: the lars algorithm with E-Net regularization leads to the highest value.

To conclude, whatever the settings, the lars algorithm with the E-Net regularization leads to the highest pROC-AUC values.

3.4.3 Mean squared errors (MSE)

Figure 3.2 and 3.3 summary the MSE values of the different methods. As data are centered and scaled, a MSE value lower than 1 means that the studied method has a prediction ability. In this case, the selected variables predict Y better than an empty set of variables. On the contrary, the studied method is not predictive if the value is above 1.

In the independent framework, all methods are predictive. The MSE values are all smaller than 0.2 for eBIC, the data-driven penalties methods, ESCV and the knockoffs method; around 0.7 for Tigress and between 0.8 and 0.9 for LinSelect. There is no significant difference between E-Net and ℓ_1 penalties or between GD and lars algorithms. Concerning the Bolasso and Stability Selection, the smallest MSE is obtained with the lars algorithm, the E-Net regularization and when samples are first generated with a value below 0.3. All values for Stability Selection are between 0.3 and 0.5; all values for Bolasso are between 0.5 and 0.7.

For the other settings, concerning cluster and scale-free-min and excepted for the data-driven penalties methods, the MSE values are smaller than 1 but close to 1 (larger than 0.92). The data-driven penalty methods are not predictive with cluster and scale-free-min (with values larger than 1) but the MSE values are smaller than 0.55 for scale-free-max. Concerning scale-free-max, the MSE values are smaller than 0.5 for ESCV, the knockoffs, Bolasso and Stability Selection when samples are first generated and with the lars algorithm and E-Net regularization. The other methods give values between 0.6 and 1, which are also predictive. A slight improvement is observed for Bolasso and Stability Selection with the lars algorithm. Otherwise, results do not depend on the regularization functions and algorithms.

To conclude, the MSE values are larger in the presence of dependency: the methods lose predictive performances when the variables are correlated, regardless of their dependence structure. The cluster and scale-free-min designs give the same comparison results: the data-driven penalties have no predictive performances and all other MSE values are smaller than 1 corresponding to predictive performances, even if values are close to 1. Concerning the scale-free-max and independent designs, variances are higher with the ℓ_1 penalty than the E-Net one. The data-driven penalty methods also present high variances and among the model selection methods, LinSelect gives the highest values. For independent design, the best method is eBIC. For scale-free-max design, the best method is the data-driven penalties with the dimension jump strategy, lars algorithm and E-Net penalty when $n = 150$ but among the variable identification methods, ESCV and the knockoffs give similar performances to the best method of the model selection

ones. The Lars algorithm and the E-Net penalties should be privileged, as well as the strategy of first generating samples.

3.4.4 Recall

Figures 3.4 and 3.5 summary the recall values of the different methods. The expected value of the recall metric is 1 meaning that all the active variables are selected.

For the independent framework, results from ESCV, the knockoffs methods and all the model selection methods apart from the LinSelect one almost equal 1 (larger than 0.94). However, LinSelect recall values are close to 0 and Tigress obtains values around 0.18. There are no significant differences between the GD and lars algorithms and the ℓ_1 and E-Net penalties for all the previously mentioned methods. The lars algorithm and the E-Net regularization function seem to be a more judicious choice for Bolasso (with value equals 0.6) and Stability Selection (with value equals 0.7). In contrast, with the ℓ_1 regularization, the values go down to 0. When the GD algorithm is used, Stability Selection obtains the best value when grids are first generated but only around 0.5 and the rest of the values are similar: around 0.3 when samples are first generated; 0.2 and 0.34 for Bolasso.

For all methods, the dependency structure decreases the recall values. More precisely, except for the data-driven penalties methods, all the others give a median lower than 0.55 for at least one design: for eBIC and LinSelect on scale-free-max and cluster (and 1 for scale-free-min); for Bolasso, Stability Selection and Tigress on the three designs; for the knockoffs on cluster and scale-free-min (and between 0.7 and 0.8 on scale-free-max); for ESCV on cluster (and 1 and between 0.7 and 0.8 for respectively scale-free-min and scale-free-max). Concerning the data-driven penalties, the slope heuristics strategy is better than the dimension jump one, having a large variability additionally: values are 1 on scale-free-min, between 0.8 and 1 on scale-free-max and around respectively 0.7 and 0.5 for slope heuristics and dimension jump on cluster.

To conclude, the results differ between the independent case and when there is a dependency structure. When variables are independent, the model selection methods recover the active variables, except LinSelect with values approaching 0. ESCV and the knockoffs also manage to recover the active variables. When a structure of dependencies exists, data-driven penalties with the slope heuristics strategy have to be privileged, especially with the GD algorithm. All model selection methods are better than all variable identification ones where ESCV is the best. Bolasso and Stability Selection are unstable with respect to the choices of regularization function, algorithm and first step generation strategy. Lastly, a slight improvement is observed for the E-Net regularization.

3.4.5 Specificity

Figures 3.6 and 3.7 summary the specificity values of the different methods. The expected value of the specificity metric is 1 meaning that all the non active variables are not selected.

For the independent setting, all the variable identification methods, eBIC and LinSelect achieve at least 0.99 or even 1 and any difference is observed between the algorithm and regularization function choices. Concerning the data-driven penalties, the strategies give varying results: the combination lars algorithm and E-Net penalty provides the highest specificity results, and the dimension jump strategy values are above the slope heuristics ones where they do not go below 0.9 except with the lars algorithm and the ℓ_1 regularization (around 0.75).

For the correlation structure settings, conclusions are the same except for slope heuristics strategies. Their values remain lower than the dimension jump ones for which the lars algorithm and the E-Net regularization get the best score around 0.9 while the others are maintained between 0.8 and 0.91. The dependency structure decreases specificity for the slope heuristics strategies and more variability.

To conclude, all methods except the data-driven penalties control the specificity well with values almost equal 1 and the specificity metric is not sensitive to the choice of both algorithm and regularization function. Hence, they are specific. Concerning the data-driven penalties, the dimension jump strategy is better than the slope heuristics one and the best combination is the lars algorithm and the E-Net regularization.

3.4.6 False discovery rate (FDR)

Figures 3.8 and 3.9 summary the FDR values of the different methods. The single method controlling the FDR is the knockoffs method and we recall that its target FDR is 0.1. Concerning the others, we expect a small value.

For the independent framework, results from all the variable identification methods equal 0 except for the knockoffs FDR values (almost 0.08, lower than the initial threshold), and ESCV obtains the lowest value. Concerning the model selection methods, FDR values of LinSelect are null while the ones of eBIC are around 0.17 and the ones of the data-driven penalties methods are larger than 0.5 excepted with E-Net penalty and the lars algorithm (with value equals 0.32 and 0.37 for respectively the slope heuristics and the dimension jump strategies). Most of the time, FDR values are smaller for the ℓ_1 than the E-Net penalty.

For the other settings, all values are almost null excepted for: (i) the data-driven penalties methods: larger than 0.66, between 0.3 and 0.55 and larger than 0.93 for respectively the cluster, the scale-free-max and the scale-free-min designs; (ii) ESCV: almost 0.2 and almost 1 for respectively the cluster and the scale-free-min designs; (iii) knockoffs: 0.08 (lower than the initial threshold) for the scale-free-max design. Thus, while ESCV is the best with the independent design, it becomes the worst method among the variable identification ones. Moreover, FDR values of ESCV have high variability. Bolasso and Stability Selection give slightly higher values with the lars algorithm and the E-Net penalty function. Moreover, Bolasso is slightly better than Stability Selection.

To conclude, among the model selection methods, LinSelect is the one to be preferred. The data-driven penalty methods always give high values, amplified when dependency exists between the

variables. Among the variable identification methods, except for ESCV, all other FDR values are smaller than 0.1, the threshold fixed for the knockoffs. Because of the variability of the results, the ℓ_1 regularization is preferred over the E-Net one and the best methods are Bolasso, Tigress and the knockoffs.

3.4.7 Impact of the high-dimension

When a method fails with respect to a metric, finding the reason is important. This subsection is devoted to investigate the impact of high dimension. Hence, if the results from a method continue to be poor with a large value of n, then it ensures that the high-dimension context is not the reason. For a neutral comparison, results of this subsection are obtained from the same datasets used in Subsections 3.4.1 to 3.4.6 that we increase to obtain sets of size $n = 300, n = 600$ and $n = 1200$. In this subsection, we only present analyses of the results but boxplots of results are provided in supplementary material available in [2]. The results are discussed per metric.

Discrimination of the active variables from the others: When n increases, values of the pROC-AUC increase too. For $n = 300$, they are all larger than 0.9, meaning that all combinations of algorithms and regularization functions successfully discriminate between active variables and non actives ones. There is no difference between the ℓ_1 and the E-Net regularization anymore, but the lars algorithm is still to be preferred: values from the GD algorithm decrease to around 0.85 in the independent and scale-free-max designs for $n = 600$ and $n = 1200$.

MSE: When n increases, approximately no change is observed for cluster and scale-free-min designs: slight decreases in values but remain close to 1. For scale-free-max and $n = 600$, all model selection methods give values between 0.2 and 0.3; the MSE values of all the Bolasso and Stability Selection methods go down to at most 0.4; while Tigress and the knockoffs remain with high values (respectively 0.8 and around 1). For independent design, prediction performances of LinSelect improve drastically since values are below 0.1 for $n = 600$; all the model selection methods give MSE values smaller than 0.3 when $n = 1200$; Tigress finishes above 0.8. Finally, concerning Bolasso and Stability Selection methods, the best combination remains the lars algorithm and when samples are first generated since the MSE values exceed 0.8 when grids are first generated for independent design and Bolasso gets the smallest values from $n = 600$ in the independent setting. Increasing the value of n changes the best method for scale-free-max which is eBIC instead of the data-driven penalties with the dimension jump strategy but does not change the best method for the three other designs (eBIC for independent and all excepted the slope heuristics strategy ones are equivalent for cluster and scale-free-min).

[2] `https://sites.google.com/view/placroix/research`

Recall: When $n = 300$, changes are observed only in the independent and the scale-free-max designs: while the knockoffs fall to 0, all others gain 0.2 in the scale-free-max, leading to values of the data-driven penalties equal 1. When $n = 600$, the knockoffs go up to 1 but only on the independent design, as LinSelect on the independent and scale-free-max designs. Concerning Bolasso and Stability Selection, values gain 0.5 and 0.4 on scale-free-max and cluster design. The best strategy is when lars algorithm, E-Net penalty and samples first generated are used. Finally, all results are 1 excepted the knockoffs on scale-free-min and from $n = 600$ to $n = 1200$. The model selection methods finish at 1 at $n = 1200$ excepted for LinSelect on cluster (with a 0.3 recall). Concerning the variable identification methods, they behave differently in designs reflecting an instability: the knockoffs remain at 0 until $n = 1200$ for cluster and scale-free-min but finish at 1 for the two others; ESCV values achieve 1 for scale-free-min and independent, 0.9 for scale-free-max with E-Net but 0.1 for ℓ_1 regularization, and 0.3 for cluster; concerning Bolasso and Stability Selection, in independent design, values are between 0.7 and 1 when samples are first generated but lower than 0.5 when grids are first generated while on dependency structures, 0.9 is obtained when grids are first generated but only 0.6 when samples are first generated. The E-Net regularization is more judicious than the ℓ_1 one. Lastly, Tigress is not sensitive, with values always lower than 0.2.

Specificity: When n increases, there is no change for the variable identification methods and eBIC with values always almost 1. However, the LinSelect performances are deteriorated: except for the scale-free-min and cluster designs where values remain at 1, the specificity results drop below 0.2 for the independent and scale-free-max designs and from $n = 600$ and are unstable with respect to the used algorithm and regularization function. The dimension jump strategy remains better than the slope heuristics one, which deteriorates faster with n. Expected for independent setting, the GD algorithm is better than the lars one.

FDR: When $n = 300$, it remains only the data-driven penalties methods with non-null FDR values: almost 0.2, larger than 0.7, between 0.5 and 0.7 and almost 1 for respectively the independent, the cluster, the scale-free-max and the scale-free-min designs. In addition and surprisingly, when $n \geq 600$, the FDR values of several methods become non-zero: the LinSelect ones climb to more than 0.8 for independent and scale-free-max designs and 0.2 for the cluster one; the eBIC ones just exceed 0.1 for cluster; and the Stability Selection ones achieve 0.5 for scale-free-min design with the GD algorithm. Bolasso, Tigress and the knockoffs have no impact with the changing values of n and they are the best methods for a FDR control point of view.

3.4.8 Results from the FRANK designs

This subsection investigates the impact of methods applied to another kind of data. Indeed, the FRANK ones are generated from a dynamic process instead of independent and Gaussian models (as the 4 other settings). The FRANK data are closer to real ones but deviate from

the statistical model's assumptions, especially the Gaussian distribution. In this subsection, we present only analyses of the results but boxplots of results are provided in supplementary material available in [3]. The results are discussed per metric.

Discrimination of the active variables from the others: Similarly with scenarios from independent and Gaussian models, the combination of the E-Net regularization with the lars algorithm achieves the highest value of pROC-AUC. However, these values equal 0.5 for FRANK-max and 0.57 for FRANK-min. All the others are smaller than 0.3. Hence, the quality of the regularization paths has clearly deteriorated on FRANK data: the studied model collection procedures do not discriminate the active variables from the others correctly.

MSE: The median MSE values are close to 1 for all methods. More precisely, MSE values are strictly smaller than 1 but strictly larger than 0.99 for all Bolasso and Stability, for Tigress and the knockoffs methods in both FRANK-max and FRANK-min. ESCV is the best method, with values between 0.97 and 0.99. It is also the case for eBIC and LinSelect but only in FRANK-min design. However, in FRANK-max one, all the model selection methods lead to a value larger than 1, meaning they are not predictive. From the data-driven penalties, values are the highest ones and always larger than 1 whatever the design. Hence, the prediction performances are deteriorated on the FRANK designs compared to the 4 other scenarios studied in this review. The model selection methods lose even their prediction performances.

Recall: Values are null for all the methods in the FRANK-min design. Concerning the FRANK-max design, ESCV achieves only 0.06, the dimension jump remains smaller than 0.08 and the slope heuristics values are below 0.24. All the other values are null. Hence, all the methods fail to select the active variables in the FRANK data.

Specificity: Values equal 1 except for ESCV (with values larger than 0.96), Bolasso with the lars algorithm, E-Net regularization and when samples are first generated (with values larger than 0.98), eBIC and LinSelect (with values larger than 0.99) and the data-driven penalties: values for dimension jump are always larger than 0.93, but the slope heuristics values can decrease until 0.7 for FRANK-max and 0.84 for FRANK-min.

FDR: In FRANK-max design, the FDR values are equal to 1 for eBIC and LinSelect, meaning that selected variables are all non active. Values are larger than 0.9, 0.88, 0.8, 0.8 and 0.6 for respectively the slope heuristics, the dimension jump, ESCV, Stability Selection and Bolasso with E-Net regularization, lars algorithm and when samples are first generated. All the other FDR values are null. In FRANK-min design, all the model selection methods achieve 1 as well as ESCV. With E-Net regularization, lars algorithm and when samples are first generated, Bolasso achieves 0.7, Stability Selection achieves 1. All the other FDR values are null. Hence,

[3] `https://sites.google.com/view/placroix/research`

compared to the 4 other scenarios, the FRANK ones lead to very high values of FDR, or even values equal to 1.

3.5 Practical recommendations

As preliminary recommendations, Tables 3.2 and 3.3 suggest that Tigress and LinSelect are both very conservative leading to an almost empty set of selected variables. Conversely, the considered data-driven penalties provide a too large set of selected variables. The E-Net regularization, the lars algorithm and the strategy of samples first generated have to be privileged to get a size of selected variables set close to the set of active variables.

3.5.1 Recommendation per method

In this following table, we summarize the best combination of the algorithm, the regularization function and eventually the strategy per method.

method	regularization function	algorithm	strategy
path:	E-Net	lars	
eBIC:	ℓ_1	lars	
data-driven penalties:	*to avoid*		
LinSelect:	indifferent	indifferent	
ESCV:	indifferent	indifferent	
The knockoffs:	indifferent	indifferent	
Tigress:	ℓ_1 (per default)	lars (per default)	
Bolasso:	E-Net	lars	samples first generated
Stability Selection:	*to avoid*		

More precisely, the data-driven penalties have to be avoided due to bad performances or instabilities in results, whatever the metric. Among the methods based on the sampling procedure, we suggest using Bolasso instead of Stability Selection. Indeed, Stability Selection gives similar or poorer results compared to Bolasso. Moreover, Stability Selection results depend most of the time on the dependence structure setting and are unstable. Lastly, the variables selection of Stability Selection is processed on samples of size $\lfloor \frac{n}{2} \rfloor$, which is even smaller than p, and so which accentuates the possible issues due to the high-dimension context. The lars algorithm and the E-Net regularization are the most present in the recommendations compared to the GD

algorithm and the ℓ_1 regularization. This combination probably gives the highest pROC-AUC values and, therefore, a better regularization path to discriminate the active variables from the others.

3.5.2 Recommendation per metric

In this subsection, we summarize best methods per metric as well as those to avoid.

Here is a summary on the choice of methods when the dependency structure is unknown to obtain performances on discrimination of the active variables from the others (high value of pROC-AUC):

pROC-AUC:	
Best regularization function:	E-Net
Best algorithm:	lars

More precisely, whatever the settings and the value of n, the lars algorithm with the E-Net regularization leads to the highest pROC-AUC values.

Here is a summary on the choice of methods when the dependency structure is unknown to obtain prediction performances (small value of MSE):

MSE:	
Best methods:	ESCV, knockoffs, LinSelect, eBIC
Best regularization function:	E-Net
Best algorithm:	lars
methods to avoid:	data-driven penalties

More precisely, ESCV and knockoffs have the smallest values of MSE and LinSelect and eBIC get also reasonably small MSE values, making them predictive methods. As for the data-driven penalties, it is on the cluster and scale-free-min designs that give an MSE greater than 1, which is undesirable. In a particular case where the dependence structure is known, the data-driven penalties method with the dimension jump strategy is the best for scale-free-max; and eBIC and LinSelect are to be considered for scale-free-min.

Here is a summary on the choice of methods when the dependency structure is unknown to obtain a support with enough active variables taking into account that the non active variables are not selected (high value of both recall and specificity):

recall and specificity:	
Best methods:	ESCV, eBIC
Best regularization function:	E-Net
Best algorithm :	GD, lars
methods to avoid:	Tigress, LinSelect, data-driven penalties

Recall and specificity metrics have to be considered together. Indeed, only controlling the recall would lead to too large support containing all the active variables but without considering the non active variables selection: active variables would be selected but mixed between many non active ones. On the opposite side, only controlling the specificity would lead to too small support avoiding certainly the non active variables selection but without taking into account the selection of active variables: the selected variables would be active, but many active variables would not be selected. Thus, only controlling specificity leads to a conservative method with a small recall, while controlling recall leads to too large selected variables set with a small specificity. Both are undesirable in practice. The ideal method is achieving the best compromise between the two metrics. Hence, Table above gives methods having reasonably high values for both metrics. More precisely, Tigress and LinSelect are set apart since their recall values are smaller than 0.5. In contrast, the data-driven penalties are set apart since their specificity values are the smallest with respect to all the others. Once these three methods are removed from the analysis, the remaining variable identification methods have the same performances in terms of specificity metric. The highest recall value allows for determining the best choice obtained with ESCV. In the same way, the remaining model selection methods have the same performances in terms of the recall metric. The highest specificity value allows for determining the best choice, which is obtained with eBIC. These values are just below those of ESCV but eBIC recall values are just above ESCV recall ones, which makes eBIC and ESCV competitive methods for both control of recall and specificity.

Here is a summary on the choice of methods when the dependency structure is unknown to obtain a support with a small number of non active variables (small value of FDR):

FDR:	
Best methods:	Bolasso, Tigress, the knockoffs, LinSelect
Best regularization function:	ℓ_1
Best algorithm:	GD, lars
methods to avoid:	ESCV, data-driven penalties

More precisely, LinSelect, Bolasso, Tigress and the knockoffs have FDR values smaller than 0.1, the threshold usually fixed in practice. ESCV and the data-driven penalties give the highest FDR values whatever the dependence structure.

3.6 Discussion

High-dimensional regression is usually used to model projects with real dependent data when the number of variables is close to or larger than the number of observations. This framework raises many methodological questions and this review aims at highlighting the method performances according to the application objectives. This work is focused on variables selection methods from regularization path constructions. We propose an evaluation of both the regularization path construction and the choice of the final selected variables. The first step is based on the least-squares penalization and the main question is the choice of the regularization function; the second step is based on either penalized criteria minimization with either asymptotic or non-asymptotic properties (model selection methods), or data sampling strategies (variable identification methods). To evaluate the different methods in a fair way, we simulated different settings, each one having its own characteristics. The independent setting is irrelevant for most of the applications since having completely independent variables is rare in practice. However, this setting is commonly used to develop the statistical methods and is a benchmark to evaluate performances of any method. The settings based on the Gaussian graphical model generating correlated variables are the most favorable case where assumptions of our model broadly hold. The FRANK setting provides a more realistic framework but deviates from the statistical model's assumptions. It is based on a dynamic process to generate a gene regulatory network. Lastly, the methods are evaluated for different performances: the ability to discriminate the active variables from the others through the pROC-AUC, the prediction quality through the MSE, the ability to recover the active variables through the recall, the ability to not select the non active variables through the specificity and the quality of the selected variables subset through the FDR.

The impact of dependency structure presence between variables is evaluated by comparing results on the independent case with those on the three other designs. Obtained results show significant degradation of performances for any metric on these three last settings. This proves that controlling the variable dependencies is an important assumption in a statistical model. Moreover, analyses and interpretations are similar on independent and scale-free-max designs: comparing methods on scale-free-max design amounts to compare them in an independent case, while those on cluster and scale-free-min designs are almost identical: comparing methods on cluster design amounts to compare them when there is only one active variable, which is equivalent to analyze the ability to select the only active variable and how many other variables (corresponding to non active ones) are selected in addition. Moreover, metrics are better controlled on scale-free-max setting than on cluster and scale-free-min ones. These observations suggest that when dependence between variables exists, the methods work better when the support is large enough, but while respecting the sparsity assumption.

Moreover, the active variable numbers average of the independent design is close to that of the cluster one. This choice allows to analyze results from the independent design as a benchmark. One of the most striking conclusion, which can already be observed on the independent setting but which is also generalized on the others, is that the best choice of a method differs according to the metric to control. Analyses are performed by studying the medians and the variability

of the results metric by metric:
- The first step of the statistical framework is to define a ranking on the variables through the subset collection construction. According to the simulation study, the combination of the lars algorithm and the E-Net regularization is the best one to discriminate the active variables for the others. Because of the truncation to compare the regularization paths in a fair way, the reference value is not 1. A perspective is to add the maximum value that can be reached by the regularization paths. We expect to obtain values far from this maximum value. Indeed, by studying the support of variables along the regularization path, some non active variables are present from the smallest supports. This phenomenon is reduced with the increase of n. Hence, the high-dimensional context impacts the variable selection procedure from the step of the regularization paths constructions.
- To define a final variable subset from the ranked variables and in a prediction consideration, although model selection procedures have theoretical guarantees, the best choices are among the variable identifications ones: ESCV and the knockoffs. eBIC and LinSelect are also competitive and lead to slightly higher values of the MSE.
- As already said, prediction differs from variable identification. To recover all active variables, all the model selection methods excepted LinSelect are good choices: they give a high recall. ESCV is the best method among the variable identification ones. However, the question is how many non active variables are also selected by these methods. This number is well controlled with all variable identification methods: there are specific. eBIC is the best method among the model selection ones. Controlling recall or specificity without a regard for the other metric is not a good idea, they have to be considered together. In this direction, ESCV and eBIC are the best choices.
- To ensure that the selected variables are active variables, a small value of FDR is expected. In this case, the best choices are the knockoffs and LinSelect, having a FDR equals to 0. Bolasso and Tigress methods get a FDR value smaller than 0.1, the common threshold used in practice. Selecting non active variables can be costly in practice: ESCV and the data-driven penalties have to be avoided.

As a general conclusion, the used strategy should be decided according to the metrics to favor. The best methods per metric give satisfactory performances: the corresponding MSE is small, both recall and specificity are high and FDR is close to 0.
- Concerning the data-driven penalties, strategies are to be avoided most of the time. One reason may be that data-driven methods are based on a heuristic whereas LinSelect, the other non-asymptotic model selection procedure, was constructed from an oracle inequality. Moreover, we also want to point out that for the data-driven calibration methods, the shape penalty and the multiplicative constant 2.5 in (3.6) are derived from other statistical frameworks. More precisely, the constant 2.5 has been fixed in a context of detection of changepoints in a signal [Lebarbier, 2005]. This value may be not adapted in a high-dimensional Gaussian linear regression leading to the absence of either the expected linear behavior between the least-squares values and the penalty shape ones, or the dimension jump. More theoretical work is needed to propose other choices which could improve their performances in high-dimensional regression.

- Among the model selection methods, the ones with asymptotic properties are commonly used in literature but this review emphasizes that the non-asymptotic ones should also be considered: especially, LinSelect results are independent of the dataset characteristics, an advantage for analyzing real data, and it is not the case for eBIC.
- This review also highlights very conservative methods that select almost no variables: this concerns LinSelect and Tigress. Hence, although they guarantee FDR values almost zero and a specificity almost equal to 1, the number of selected variables is very low and far from the number of active variables. It is consistent with the practical recommendations: both have to be avoided to control a trade-off between recall and specificity. To select more variables, accepting a non-zero but controlled FDR and a specificity close to 1 is suggested and guaranteed by Bolasso and the knockoffs method. Broadly speaking, all the methods are conservative except for the data-driven penalties. The last ones have to be avoided for all metrics: too much non active variables are usually selected compared to the active variables. ESCV is the less conservative method but has to be avoided to control FDR meaning that among the selected variables, some of them are non actives.
- As for the sampling strategies, Bolasso should be preferred to Stability Selection since estimation from Stability Selection is more tricky due to the subsamples using instead of resamples one leading to a greater instability. Generating samples first rather than grids significantly reduces computational time.
- Concerning the algorithms, the lars one has to be preferred compared to the GD one, whatever the metric; concerning the regularization function, E-Net achieves the best results excepted for the FDR metric where ℓ_1 leads to a smaller value. Most of the time, the E-Net regularization leads to best performances on the independent design suggesting that the ℓ_1 regularization procedure sets the coefficients to zero too suddenly and adding the regularization by the Ridge improves results even in the absence of dependency structure.

The choice of parameters within a method does not have the same importance as the choice of a method to answer a problem. A perspective of this work would be to study the impact of the parameters variation. Moreover, a statistical framework always requires some hypotheses. For the high-dimensional Gaussian linear regression, data are assumed to be distributed according to a Gaussian distribution, observations are supposed to be independent, variable dependencies are well controlled and the set of active variables is supposed to be small. Real datasets generally do not verify all these hypotheses. The relaxation of each of these hypotheses can be studied in future works. For instance, the sparsity assumption can be easily relaxed by considering larger sizes of active variables subsets under the independent design. In this work, we just investigate:
- the impact of the high-dimensional issue by evaluating metrics on augmented data given samples for $n = 300, n = 600$ and $n = 1200$. If the results from a method continue to be poor with a large value of n, then it ensures that the high-dimension context is not the reason. When $n = 1200$, we can consider that there is no longer any problem of high-dimension. It is the case for the data-driven penalties; for Tigress where values continue to be high for the MSE and small for the Recall and the knockoffs method. On the one hand, performances of model selection procedures for recall control; eBIC and the variable identification methods for specificity one; LinSelect, Bolasso, Tigress and the knockoffs for FDR one already exist when

n is large enough. Thus, their performances are not due to the context of the high-dimension. On the other hand, MSE values of the model selection procedures significantly decrease when n increases and eBIC becomes the best method whatever the metric. It is expected since its theoretical guarantees are based on an asymptotic point of view, achieving them in practice with n large enough. However, FDR values drastically increase with n for LinSelect, eBIC and Stability Selection; the specificity value increases too for LinSelect; and the knockoffs, ESCV, Bolasso and Stability Selection become unstable in terms of the recall metric. All the last results are surprising.
- the importance of the Gaussian assumption by evaluating metrics on the FRANK data. Indeed, a complementary study shows that the distribution of FRANK data is far from the Gaussian one. Since our study's conclusions are the deterioration of values for all metrics, the Gaussian distribution seems to be an important assumption for the considered statistical model. More precisely, the pROC-AUC values decrease significantly compared to the 4 other scenarios. All the model selection procedures are not predictive since the MSE values are larger than 1. The Recall values do not exceed 0.24 and some FDR values achieve 1. As for specificity, values remain high. Hence, the selected variables sets contain only or almost only non active variables. Conclusions could indicate the importance of data transformations and pre-processing steps if a real dataset's Gaussian assumption is not satisfied (see [Liu et al., 2012]). Note that in a complementary study, the *shrinkage* transformation, which is available in the R function *huge.generator* from the R package *huge*, is tested on FRANK-max and FRANK-min to replace the classical normalization per variable. The MSE values are slightly improved but all the other metrics are slightly deteriorated (boxplots are provided in supplementary material available in [4]).

Supplementary data

The scripts are available from [5].
Boxplots of results in Subsection 3.4.7 and 3.4.8 are provided in supplementary material available in [6].

3.7 Appendix: Boxplots for $n = 150$ per metric

In this appendix, boxplots of results from scenarios independent, cluster, scale-free-max and scale-free-min when $n = 150$ are provided. They are arranged by metric. For boxplots of results when $n = 300, 600, 1200$ as well as for boxplots of results from the FRANK scenarios, the reader is invited to consult the supplementary material available in [7].

[4] `https://sites.google.com/view/placroix/research`
[5] `https://forgemia.inra.fr/GNet/high-dimensional_regression_comparison`
[6] `https://sites.google.com/view/placroix/research`
[7] `https://sites.google.com/view/placroix/research`

The *glm*, *enet* and *lasso* denote respectively the gradient descent algorithm, the Elastic-Net regularization function and the Lasso regularization function. The slope heuristics method and the dimension jump are named respectively by *slope* and *jump*. Bolasso and Stability Selection are named respectively by *b* and *ss*. Lastly, *grid* and *sub* denote respectively the strategy when grids are first generated and the strategy when samples are first generated.

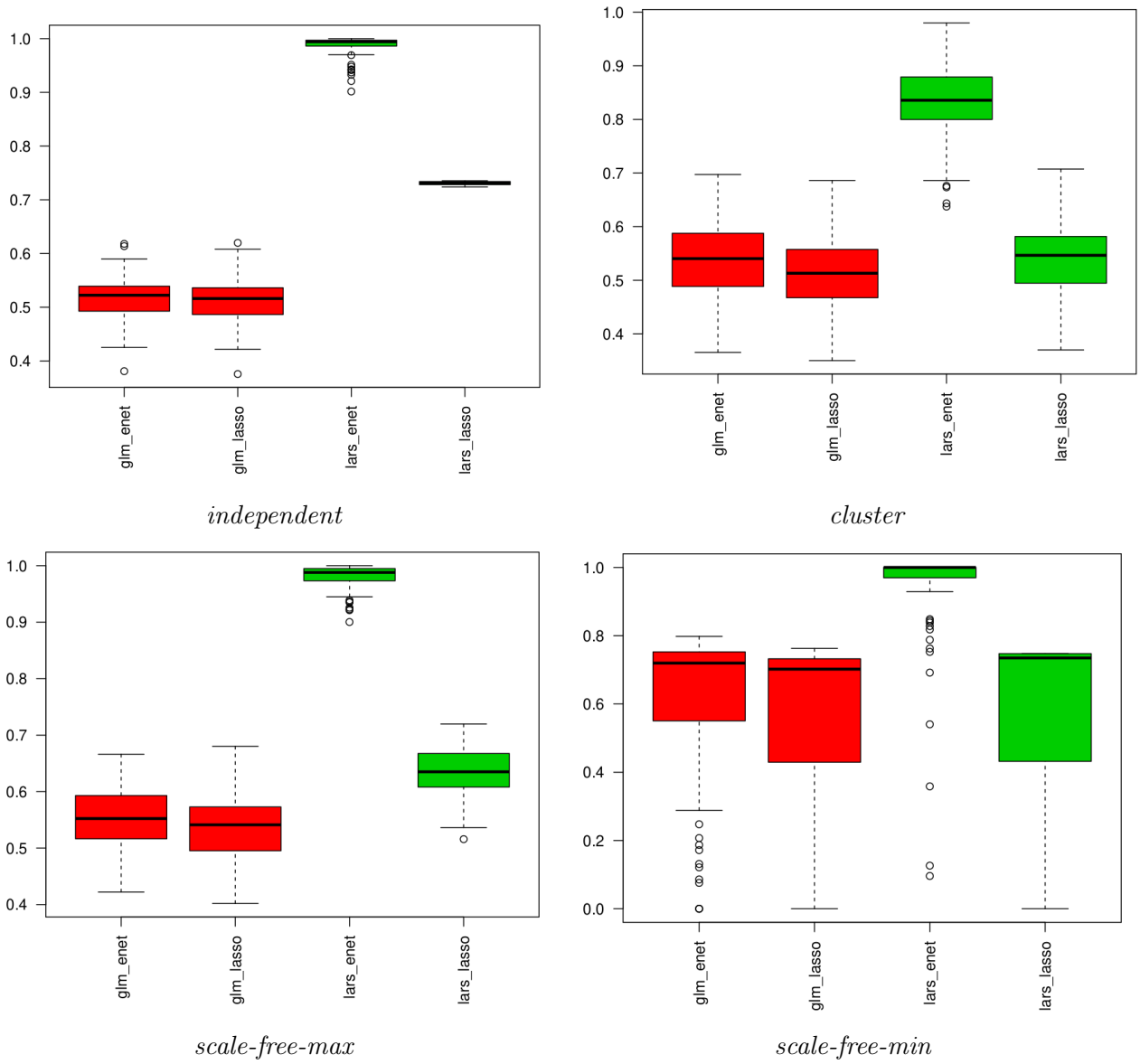


Figure 3.1: Boxplots of the pROC-AUC values for $n = 150$.

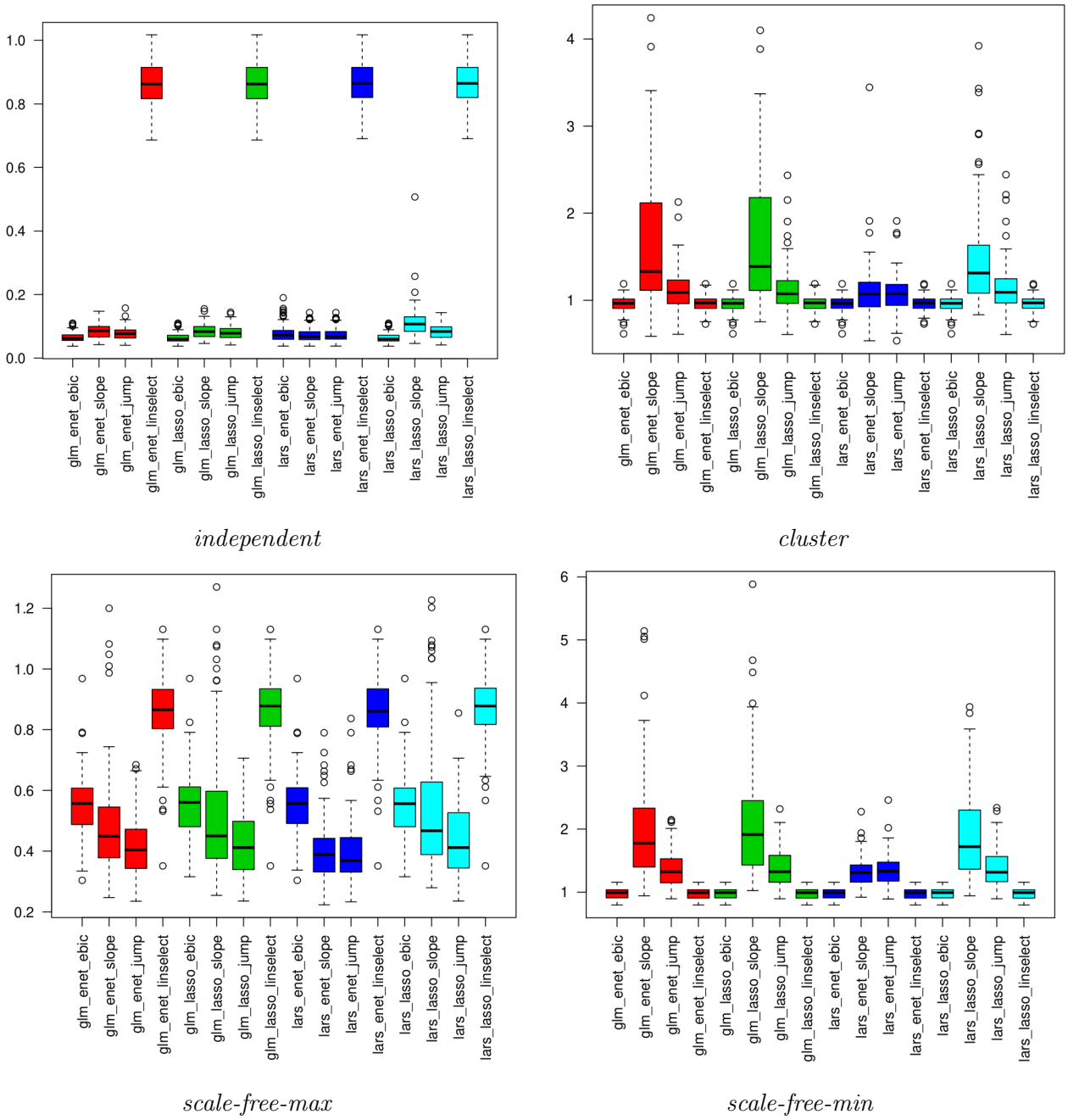


Figure 3.2: Boxplots of the MSE values for model selection procedures and for $n = 150$.

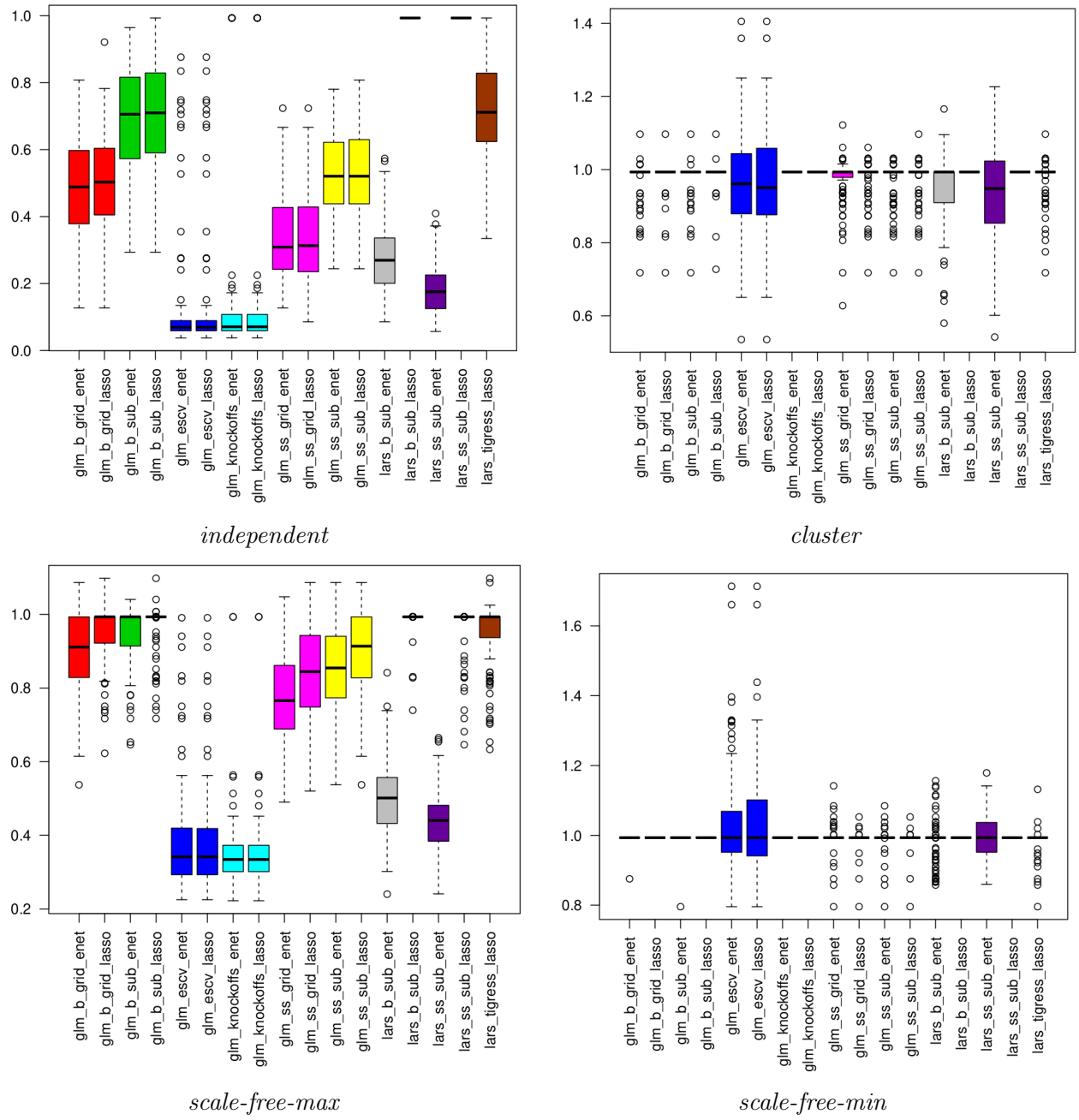


Figure 3.3: Boxplots of the MSE values for variable identification procedures and for $n = 150$.

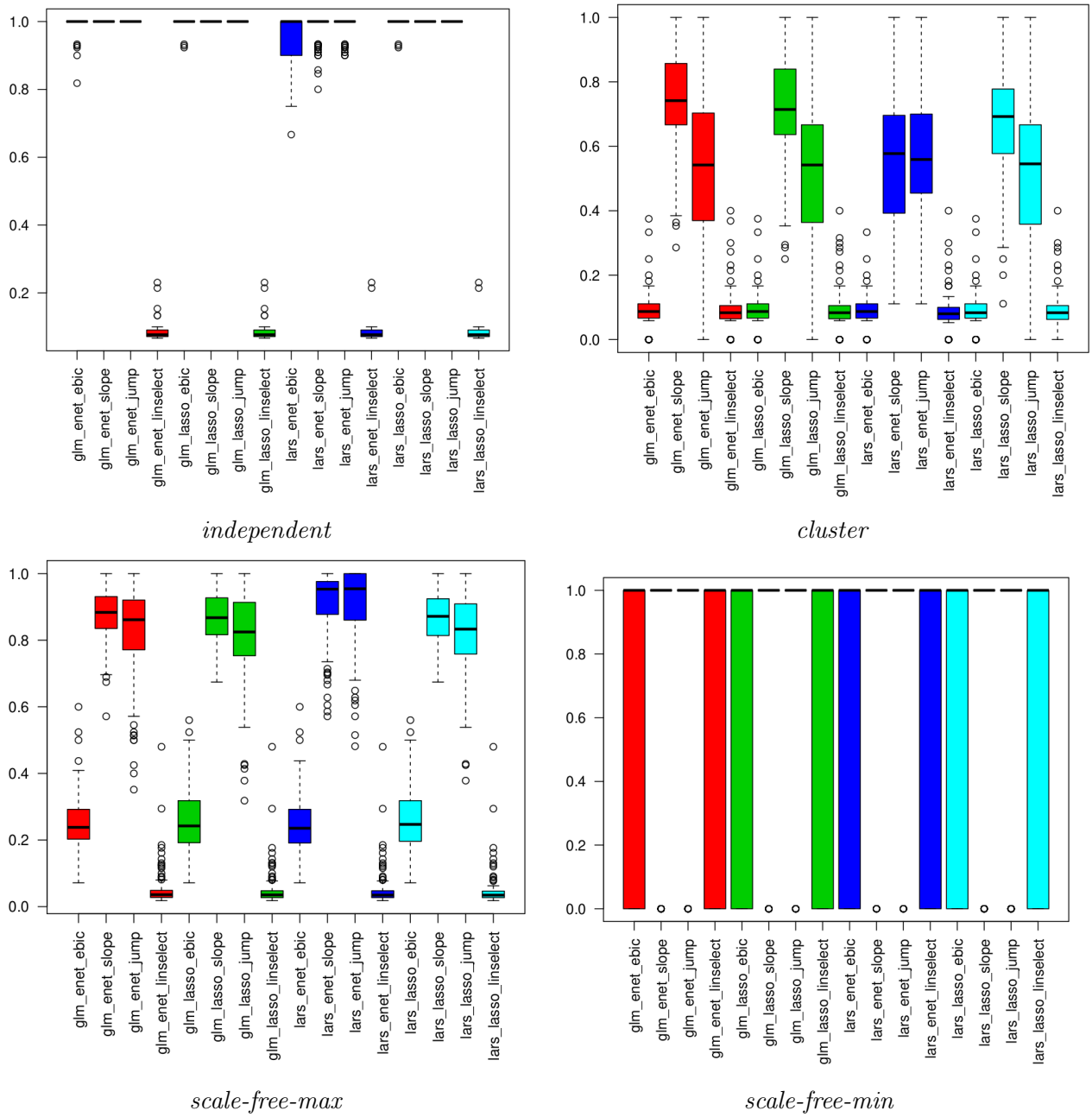


Figure 3.4: Boxplots of the recall values for model selection procedures and for $n = 150$.

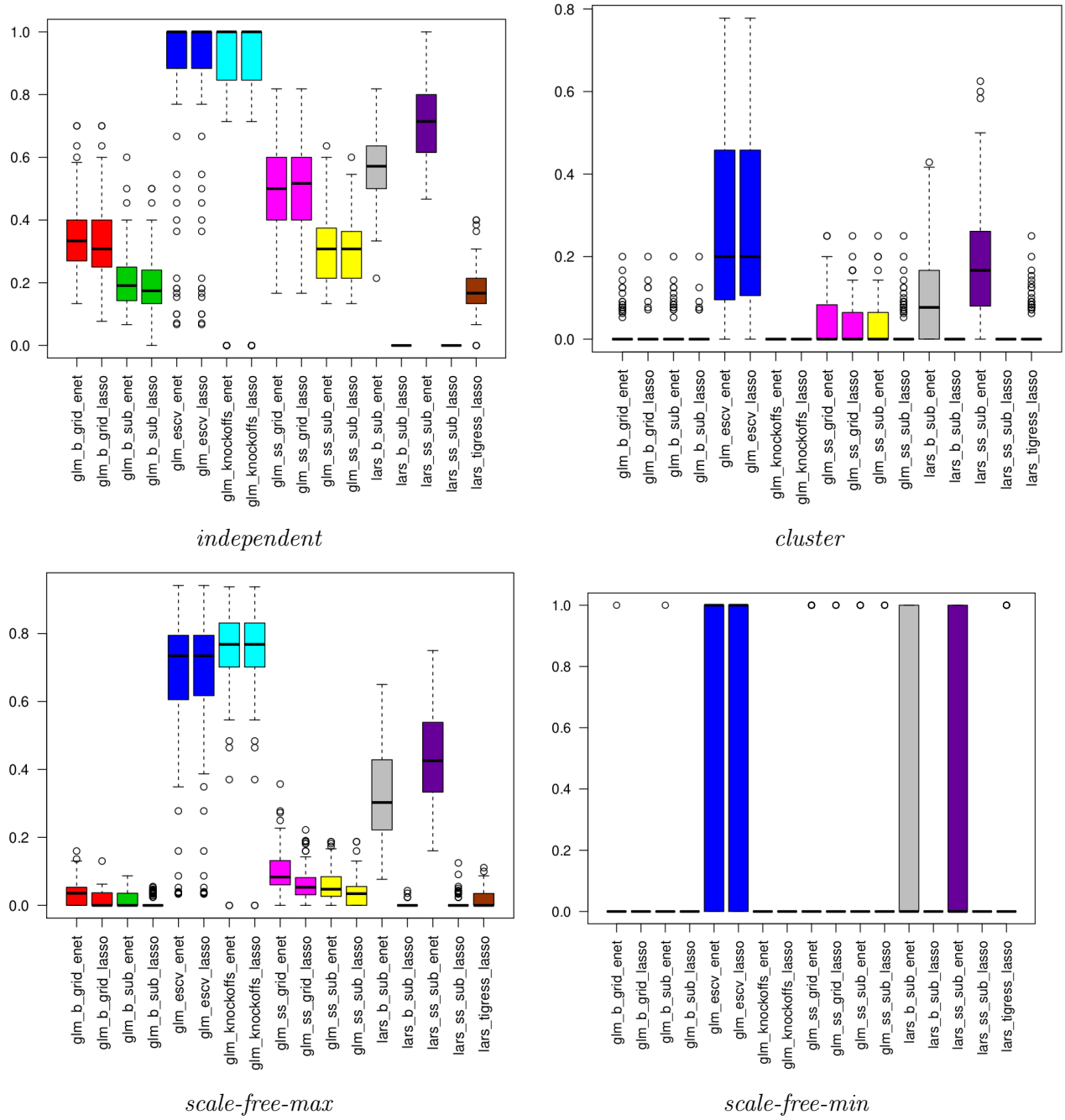


Figure 3.5: Boxplots of the recall values for variable identification procedures and for $n = 150$.

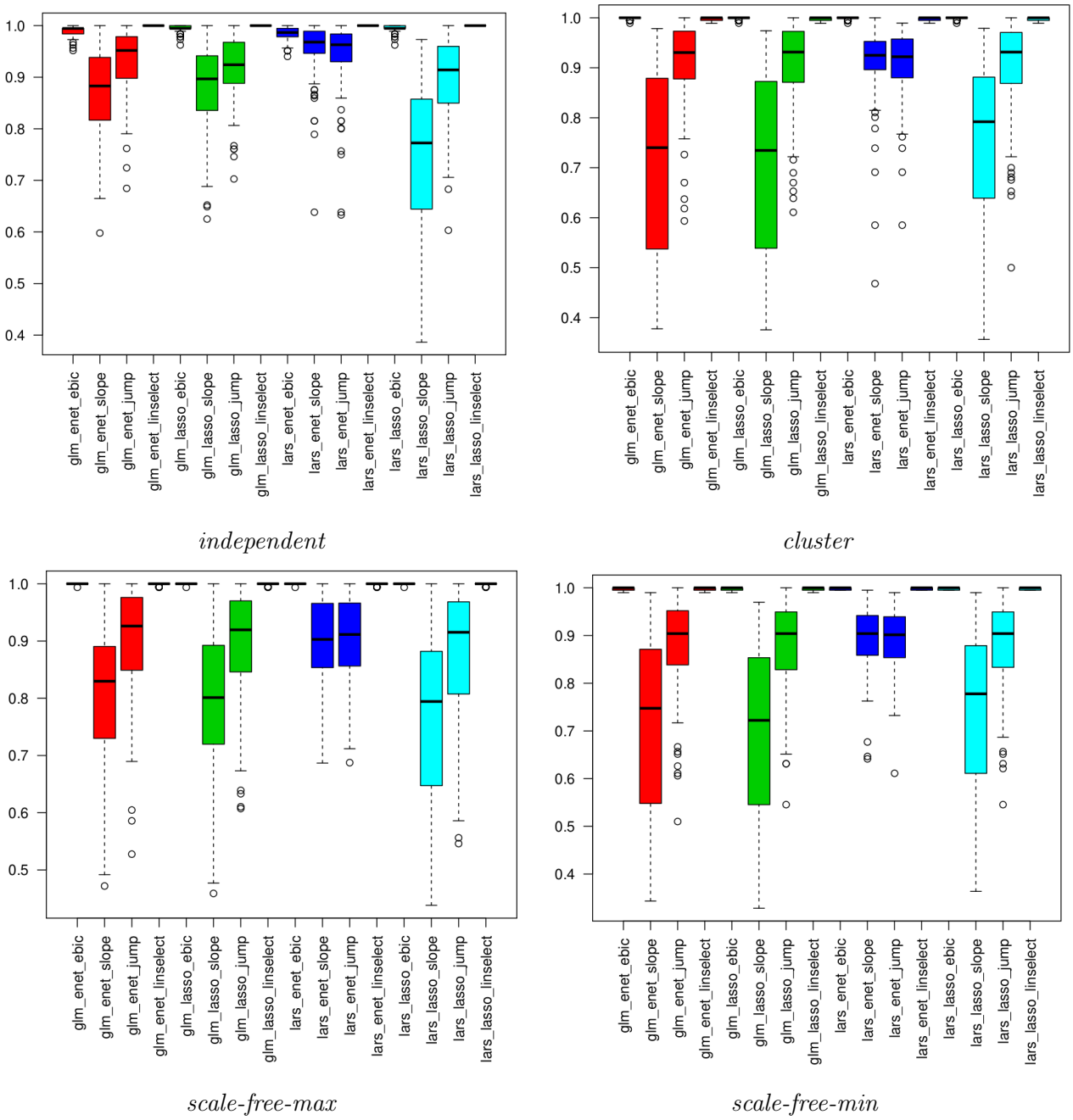


Figure 3.6: Boxplots of the specificity values for model selection procedures and for $n = 150$.

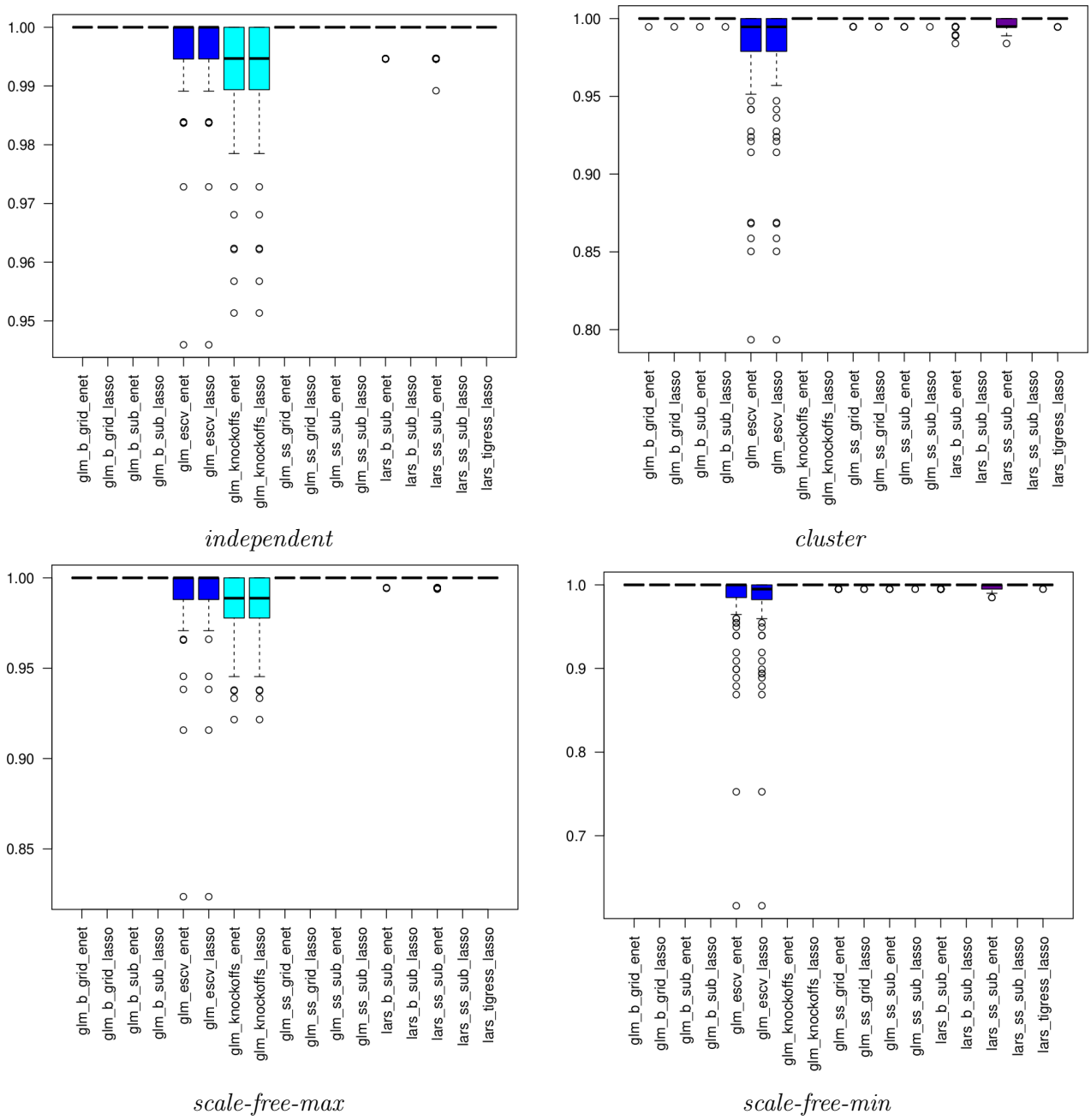


Figure 3.7: Boxplots of the specificity values for variable identification procedures and for $n = 150$.

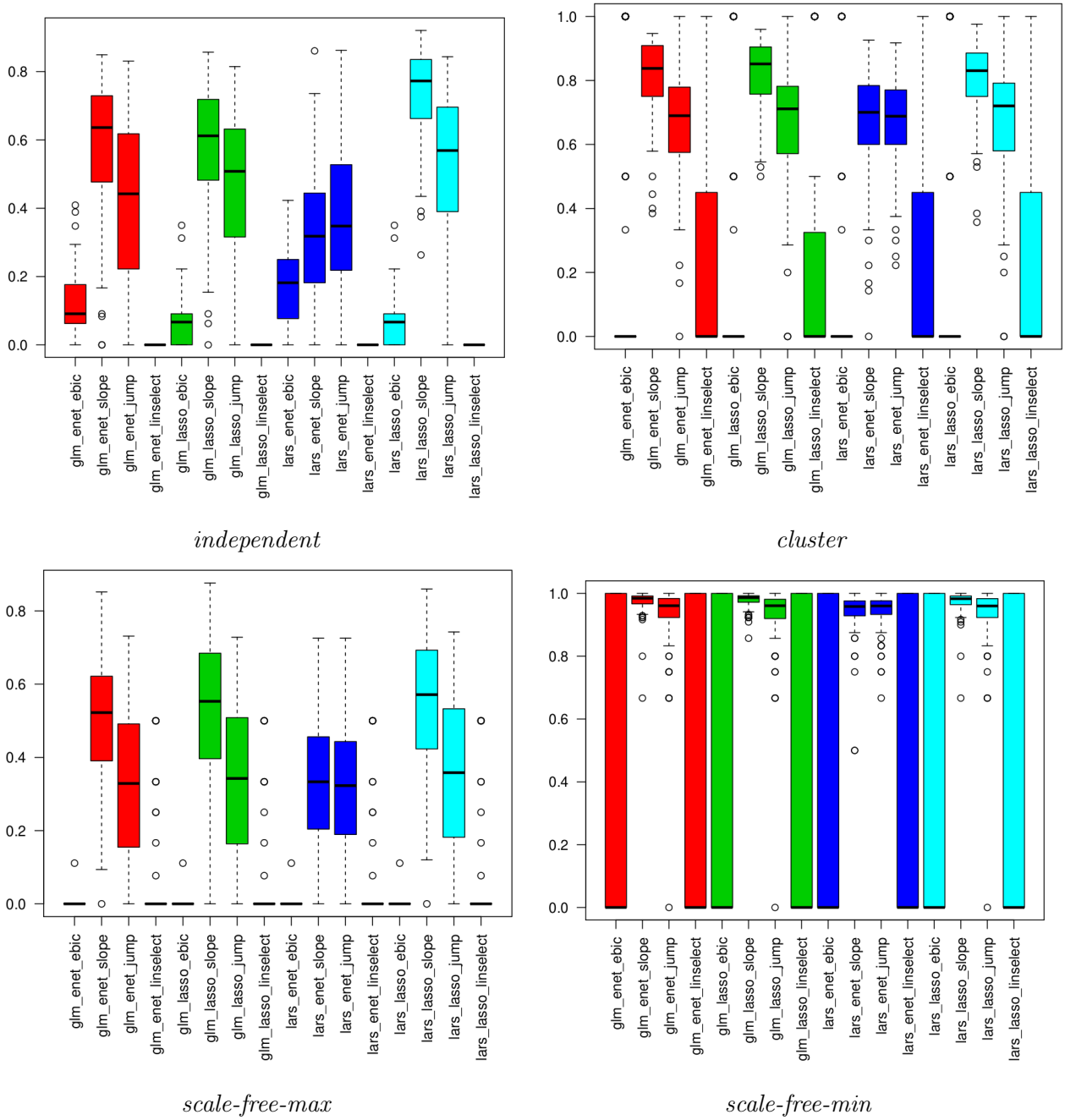


Figure 3.8: Boxplots of the FDR values for model selection procedures and for $n = 150$.

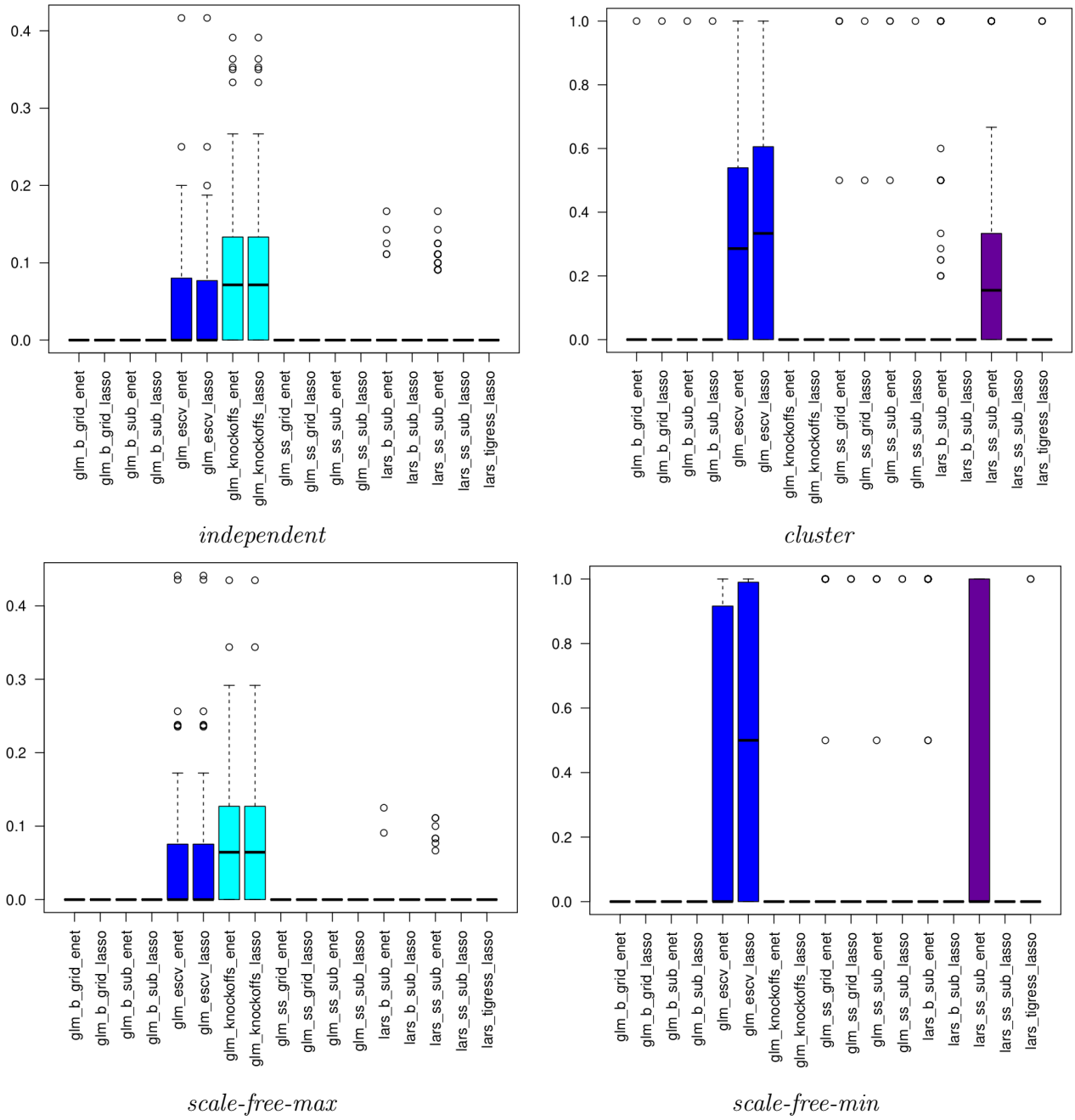


Figure 3.9: Boxplots of the FDR values for variable identification procedures and for $n = 150$.

Chapter 4

Trade-off between prediction and FDR for high-dimensional Gaussian model selection

Abstract

The studied model is the Gaussian linear regression with a high-dimensional context. We focus on the ordered variable selection procedure which is commonly studied independently by the estimation point of view through the minimization of the predictive risk and the multiple testing points of view through controlling the false discovery rate. We propose to consider both cost functions simultaneously: the developed strategy allows to set up of a data-driven penalty calibration given a trade-off between ensuring predictive performances and avoiding the selection of non active variables. We propose a new data-dependent calibration of the hyperparameter K appearing in the penalty function during the penalized least-squares minimization in the model selection procedure. Firstly, we obtain non-asymptotic theoretical controls of the False Discovery Rate with respect to K. Secondly, theoretical results allow calibrating the hyperparameter K based on completely data-driven terms to get a trade-off between the predictive risk and the false discovery rate controls. This calibration relies on an extensive simulation study.

Keywords: Model selection, FDR, High-dimension, Penalty functions calibration

The R scripts of this Chapter 4 are available at `https://github.com/PerrineLacroix/Trade_off_FDR_PR`.

Contents

4.1 Introduction.

4.1.1 Problematic.

We consider the following high-dimensional Gaussian linear regression model:

$$Y = X\beta^* + \varepsilon. \tag{4.1}$$

The random response vector $Y = (y_i)_{\{1\leq i\leq n\}} \in \mathbb{R}^n$ is regressed on p deterministic vectors: $X_1 = (x_{1i})_{\{1\leq i\leq n\}}, \cdots, X_p = (x_{pi})_{\{1\leq i\leq n\}}$. We denote by $X = (X_1, \cdots, X_p)$ the design matrix of size $n \times p$. The noise $\varepsilon = (\varepsilon_i)_{\{1\leq i\leq n\}}$ is assumed to be Gaussian: $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ with $\sigma^2 > 0$. We assume that p is of the order or larger than n. In this context of high-dimensional framework, additional assumptions of regularity are required and we assume that β^* is sparse, meaning that only a few coefficients are non-zero. Moreover, to avoid the ultra high-dimensional setting [Verzelen et al., 2012], we consider couples (n, p) such that $2k\log(\frac{p}{k}) < n$ is satisfied where k denotes the number of non-zero coefficients in β^* (see [Giraud et al., 2012]).
A variable X_j corresponding to a non-zero coefficient β_j^* is called active variable. Otherwise the variable is said to be non active. The goal is to identify a set of active variables among $X_1, \cdots, X_p$, by estimating β^* with accurately prediction performances while controlling the selection of non active variables.
In the context of high-dimension, we process the variable selection procedure.
To give a sense of the variable selection procedure, cost functions to control have to be defined. To the best of our knowledge, most procedures dealing with variables selection in Gaussian linear regression can be classified into two separated groups. The first one focuses on the estimation problem where the goal is to predict Y accurately through the estimation of $X\beta^*$ or β^*. The cost function to control is usually the *predictive risk*. The second one focuses on the multiple testing problem where the goal is to control the number of selected non active variables. The cost function to control is usually the *false discovery rate*. In the following subsection, we present a non exhaustive review of these two procedure groups.

4.1.2 Related works.

The penalized methods to control the predictive risk (PR). The first group contains the penalized methods. The main objective is correctly predicting Y by controlling the predictive risk. The penalization procedure balances between the goodness of fit and the number of variables included in the selected subsets of variables: the smaller the penalty, the higher the number of selected variables but the better the fitting to the data. The main challenge lies in the calibration of hyperparameters to adjust the trade-off. The most popular method in high-dimension is given by [Tibshirani, 1996]. They estimate β^* by minimizing the Lasso criterion:

$$\hat{\beta}_\lambda = \underset{\beta\in\mathbb{R}^p}{\arg\min}\Big\{||Y - X\beta||_2^2 + \lambda|\beta|_1\Big\} \tag{4.2}$$

where $|\beta|_1$ is the ℓ_1-norm on $\mathbb{R}^p$ and $||.||_2$ designs the usual euclidean norm of a vector of $\mathbb{R}^n$. The main challenge is calibrating the hyperparameter $\lambda > 0$ that makes the balance between adjustment and sparsity. On the one hand, one can fix λ directly. If λ is chosen to be proportional to $\sigma\sqrt{\frac{\log(p)}{n}}$, then theoretical controls of the predictive risk are satisfying [Bunea et al., 2007b,Bunea et al., 2007a]. On the other hand, one can not fix λ directly. In this direction, an alternative to not choosing λ is to solve the Lasso equation on several replicates of the initial data set (by using subsamples [Meinshausen and Bühlmann, 2010] or resamples [Bach, 2008]). The selected variables appear in most of sets of selected variables from all the replicates. The two major gains are the robustness and the stability on λ: whatever the value of λ in (4.2) and whatever a small variation in the data set, the set of selected variables is almost unchanged. Another approach consists in solving (4.2) on a relevant grid Λ of λ. This is the model selection procedure where the model collection comes from a regularized least-squares minimization procedure (like the Lasso equation minimization). In this chapter, we consider this procedure. It is based on three steps. The first step consists in solving (4.2) on Λ. Each $\lambda \in \Lambda$ provides an estimator and its support constitutes a variable subset. So, a collection $\mathcal{M}$ containing relevant subsets of variables with a wide range of sizes is generated. In the second step, the least-squares estimators onto each variable subset of $\mathcal{M}$ are calculated leading to a collection of estimators

$$\hat{\beta}_m := \underset{\{\beta, X\beta \in m\}}{\arg\min} \left\{ ||Y - X\beta||_2^2 \right\}.$$

Lastly, the third step consists in the selection of the best set of the collection. This last step is performed by solving the penalized residual least-squares minimization:

$$\hat{m} = \underset{m \in \mathcal{M}}{\arg\min} \Big\{ ||Y - X\hat{\beta}_m||_2^2 + \text{pen}(D_m) \Big\}, \tag{4.3}$$

where D_m is the dimension of the model m and the function pen is a penalty function increasing with D_m. Selecting $\hat{m}$ by minimizing (4.3) is equivalent to select $\hat{\lambda}$ in the Lasso equation, meaning that $\hat{m}$ is the model realizing the best compromise between the goodness of fit and sparsity within the available model collection. Among the most famous methods for estimator selection, we can cite $V-$fold cross-validation [Geisser, 1975, Arlot and Celisse, 2010], AIC [Akaike, 1973], CP-Mallows [Mallows, 2000], BIC [Schwarz et al., 1978] and eBIC [Chen and Chen, 2008]. For these penalty functions, theoretical guarantees are obtained when σ^2 is known and when the sample size n tends to infinity. In our case, n is fixed, relatively small, and possibly smaller than the dimension p. A non-asymptotic point of view is preferable to get properties for all values of (n, p). In this direction, [Birgé and Massart, 2007] propose some penalty functions depending on the collection complexity such that $\hat{m}$ guarantees non-asymptotic optimal control of the predictive risk. For instance, if the model collection is nested with a known variance, $\text{pen}(D_m) = 2\sigma^2 D_m$ allows to achieve an optimal non-asymptotic control of the predictive risk [Akaike, 1973]. If the model collection is fixed and large (for instance with an exponential growth with D_m) and if the variance is unknown, this optimal control is obtained with the data-driven penalties [Birgé and Massart, 2007, Baudry et al., 2012]. Lastly, the LinSelect penalty [Baraud et al., 2009, Giraud et al., 2012] guarantees the predictive risk optimal control when the model collection is random and large and with an unknown variance.

The multiple testing methods to control the false discovery rate (FDR). The second group of variable selection for Gaussian linear regression contains the multiple testing procedure. The p tests $H_0 = \{\beta_j^* = 0\}$ under $H_1 = \{\beta_j^* \neq 0\}$ are processed simultaneously given a list of p p-values. The smaller the p-value is, the higher the confidence to select X_j is. Hence, to be close to the set of active variables, the selected variables are those with small p-values. The challenge is to find the threshold q on p-values such that the p-values smaller than q give the set of selected variables. Firstly, the threshold has been estimated to control the familywise error rate which is the probability of selecting at least one non active variable [Bonferroni, 1936, Simes, 1986]. However, that leads to a conservative procedure leading to a tiny set of selected variables. A relaxation consists in the introduction of the global criterion FDR. The authors of [Benjamini and Hochberg, 1995] first provide a threshold on p-values to control the FDR at a level α assuming some assumptions: independence of the p-values and not in the high-dimensional context. These hypotheses were then relaxed in [Benjamini and Yekutieli, 2001, Storey et al., 2004, Romano et al., 2008] or in [Leung and Sun, 2021].
The authors of [Barber and Candès, 2015] propose the knockoff filter method to control the FDR. They are inspired by multiple testing through test statistics, but the procedure differs. Indeed, their idea is to create a copy $\tilde{X}_j$ of each variable, called knockoff, such that the correlations between copies are identical to those of the original variables but copies are now independent of Y. Hence, all copies are non active variables with respect to Y, so the main idea is that active variables have to be selected before their knockoff. The Lasso criterion minimizing (4.2) is applied with $X = \big(X_1, \cdots, X_p, \tilde{X}_1, \cdots, \tilde{X}_p\big)$. Then, for each variable, a statistic is computed. This statistic tends to be large if the associated variable is active and small otherwise. The threshold on these values balancing active and non active variables is computed to control the FDR.

PR or FDR is not sufficient to recover the set of active variables. Commonly, the two groups of variable selection mentioned in previous paragraphs are studied independently in the literature and yield different sets of variables. For a PR control, selected variables are predictive, meaning that new observations of them allow predicting a new observation of Y. However, some selected predictive variables can be non active. Conversely, for a FDR control, the focus is to reduce significantly the proportion of non active variables among the selected variables. Consequently, procedures tend to be conservative, so some active variables cannot be in the selected variables set, which can lead to poor prediction performances. Ideally, the goal is to select all but only active variables. This is very challenging in a high-dimensional context.

The simultaneous control of several cost functions. Recent works have been proposed to combine prediction and FDR approaches. For instance, [Zhou, 2009] use a multi-step algorithm combining the Lasso criterion and a threshold procedure on the Lasso estimator. In addition to prediction performances, a consistency property on the selected variables is satisfied. More precisely, they bound the ℓ_2-loss between β^* and the proposed estimator under some conditions on the variables $X_1, \cdots, X_p$. Another idea is the post-selection inference where the principle is to test the relevance of each selected variable [Berk et al., 2013, Lee et al., 2016].

Valid confidence intervals are provided from conditional hypothesis tests for each model of the model collection in addition to the PR control given by the regularization procedure. Their work has been generalized by [Hyun et al., 2018, Chen et al., 2021, Duy and Takeuchi, 2021] and a review can be found in [Zhang et al., 2022]. In a completely different direction, [Genovese and Wasserman, 2002, Genovese et al., 2004] propose to control the FNR (false negative rate) in addition to the FDR. A good FNR control ensures that most active variables are selected, while a good FDR control ensures that not too many non active variables are selected. The trade-off is processed by minimizing a weighted sum of both criteria. Hence, controlling both rates can provide a set of variables close to the set of active variables. However, optimal control of both FDR and FNR simultaneously cannot be achieved since an improvement in the control of one causes a deterioration in the other. Only a compromise between them can be considered. In the same way, combining estimation and selection leads to control of two criteria simultaneously: the first to limit the non active variables selection (the FDR) and the second to select variables with prediction performances (the PR). Unlike the FNR whose minimization competes with that of the FDR, the two criteria, PR and FDR, are not in opposite directions since they are based on two different statistical purposes: the inference for the PR and the decision theory for the FDR. Firstly, for the sparse high-dimensional multivariate standard mean regression with known variance, [Abramovich et al., 2006] propose a penalty function in the model selection procedure built from the multiple testing procedure of [Benjamini and Hochberg, 1995]. They get simultaneously sharp asymptotic minimality of the FDR and the PR. Then, [Bogdan et al., 2013] study the FDR of the Sorted ℓ_1 penalized estimator (SLOPE) in the sparse high-dimensional Gaussian linear regression with a known variance. The SLOPE estimator is the minimizer of the Lasso criterion (4.2) where λ is replaced by a p-vector built from the multiple testing procedure of [Benjamini and Hochberg, 1995]. When the variables $X_1, \cdots, X_p$ are orthogonal, they get the same non-asymptotic control of the FDR as in [Benjamini and Hochberg, 1995] additionally to the asymptotic minimax convergence rate of the PR. This asymptotic convergence of the FDR has been generalized under a wide range of high dimensional generalized linear models, for instance, for a random design in [Kos and Bogdan, 2020].

4.1.3 Main contributions.

Previous works on variable selection procedures usually control only one cost function. We propose to task with both PR and FDR cost functions simultaneously. The goal is to estimate β^* with accurate prediction performances while controlling the number of non active selected variables. We differ from the previous works taking into account PR and FDR. Indeed, our controls are non-asymptotic, whereas those controlling both PR and FDR simultaneously are asymptotic. Moreover, in previous works, the FNR is the common criterion to oppose to the FDR for a non-asymptotic point of view. However, FDR and FNR are on the opposite side: improving the control of one degrades that of the other. Conversely, PR and FDR have different objectives and are thought to be not in competition. On the one hand, controlling the FDR tends to give a set of variables included in the set of active variables. On the other hand,

controlling the PR tends to provide a set of variables containing the set of active variables. Moreover, unlike the FNR, the PR has information about the available dataset: adding a variable in the selected variables subset drastically reduces the PR in the learning phase. In contrast, in the over-fitting phase, the PR increase is proportional to the noise σ^2. Hence, controlling both PR and FDR would control the number of non active selected variables and would select predictive variables, ensuring a gain in prediction performances.
To obtain control of both FDR and PR criteria, we propose to modify the penalty function in (4.3). For this purpose, we assume being in a known variance σ^2 and in an ordered variable selection framework: the most relevance choice of one variable to explain Y in the sense of (4.1) is X_1, the most relevance choice of a set of two variables to explain Y is the set composed of X_1 and X_2 and so on. So, according to this order, the active variables are the first variables. A natural model collection is the one that contains nested models respecting the order on the variables. According to [Birgé and Massart, 2001], the penalty to choose in this context is:

$$\mathrm{pen}(D_m) = K\sigma^2 D_m, \quad \forall m \in \mathcal{M}. \tag{4.4}$$

This penalty provides a non-asymptotic control of the PR if K is fixed strictly larger than 1. The only remaining question is how to calibrate $K > 1$ to maintain this PR control while also ensuring control of the FDR. The study of the FDR with respect to $K > 0$ is the key idea to find a trade-off between the PR and the FDR in model selection. As the FDR is unavailable in practice, we propose to estimate the hyperparameter $K > 1$ by using lower and upper bounds of the FDR.

Theoretical bounds of the FDR in model selection: Although the model selection procedure is built for a PR control, we obtain theoretical and non-asymptotic explicit lower and upper bounds of the FDR with respect to $K > 0$ in ordered variables selection. These bounds are easily implementable and are based on a fully explicit form since they only involve some evaluations of cumulative functions of the standard Gaussian and some chi-squared variables. Moreover, they do not depend on the data but only on the model parameters. Whatever the noise amplitude, FDR is always strictly positive. When K tends to infinity, a convergence of the FDR to 0 is obtained with an exponential rate: a low value of the FDR is already ensured for not too high K values.

Calibration of the hyperparameter K: The obtained theoretical bounds depend on the model parameters β^* and σ^2. From a large simulation study, we propose to estimate σ^2 by the slope heuristics principle and the estimator $\hat{\beta}_{\hat{m}(4)}$, defined in (4.6) and (4.7), to replace β^* in the FDR bounds. These estimators lead to completely data-dependent terms. Our proposed data-driven algorithm calibrates the hyperparameter K appearing in the penalty function. It is calibrated to be the smallest one belonging to the intersection of the interval of smallest values of estimated risk and the interval of strictly positive values smaller than a given threshold of the estimated upper bound of the FDR. A trade-off between PR and FDR is provided. Our algorithm is validated on an extensive simulation study with different model parameters.

4.1.4 Plan of the chapter.

The rest of the chapter is organized as follows. Section 4.2 introduces the model and the notation. In Section 4.3, we propose the theoretical results. In Section 4.4, we present the hyperparameter calibration of the penalty function to ensure a trade-off between FDR and PR controls. Lastly, Section 4.5 contains proofs of results and Section 4.6 is a conclusion.

4.2 Model and notation.

Let us consider the Gaussian linear regression model (4.1). For the statement of theoretical results, σ^2 is supposed to be known; for applications, σ^2 is supposed to be unknown. We define $q = \min(n, p)$ and we assume that $(X_1, \cdots, X_q)$ is a family of linearly independent vectors. We consider the deterministic and nested model collection of linear spaces:

$$\mathcal{M} = \Big\{ m_0 = \{0\}, m_1 = \mathrm{Span}(X_1), m_2 = \mathrm{Span}(X_1, X_2), \cdots, m_q = \mathrm{Span}(X_1, X_2, \cdots, X_q) \Big\}. \tag{4.5}$$

For each $m \in \mathcal{M}$, D_m is the dimension of m and $\hat{\beta}_m$ is the least-square estimator onto m:

$$\hat{\beta}_m = \underset{\{\beta, X\beta \in m\}}{\arg\min} \Big\{ ||Y - X\beta||_2^2 \Big\}. \tag{4.6}$$

With the definition of q and properties on the family $(X_1, \cdots, X_q)$, $\hat{\beta}_m$ is unique for each $m \in \mathcal{M}$. We assume that the true model $\mathrm{Span}\big(X_j,\ j \text{ s.t. } \beta_j^* \neq 0\big)$, that we denote by m^*, belongs to $\mathcal{M}$. For all $K > 0$, we define the function crit_K on $\mathcal{M}$ as:

$$\mathrm{crit}_K(m) = ||Y - X\hat{\beta}_m||_2^2 + K\sigma^2 D_m,$$

and the selected model $\hat{m}(K)$ by:

$$\hat{m}(K) = \underset{m \in \mathcal{M}}{\arg\min} \Big\{ \mathrm{crit}_K(m) \Big\}. \tag{4.7}$$

For each $m \in \mathcal{M}$, we define $\mathrm{PR}(m)$ the predictive risk associated to the model m:

$$\mathrm{PR}(m) = \mathbb{E}\Big[||Y - X\hat{\beta}_m||_2^2 \Big] \tag{4.8}$$

where $\mathbb{E}$ designs the expectation under the distribution of Y satisfying (4.1). We define $FP(m)$ the number of variables contained in m but not in m^*, the false discovery proportion by:

$$\mathrm{FDP}(m) = \frac{\mathrm{FP}(m)}{D_m \vee 1}$$

where for all real values a and b, $a \vee b = \max(a, b)$; and the false discovery rate by:

$$\mathrm{FDR}(m) = \mathbb{E}\Big[\mathrm{FDP}(m) \Big].$$

Finally, $\langle ., . \rangle$ designs the canonical scalar product in $\mathbb{R}^n$, $\Pi_{\mathcal{X}}$ denotes the orthogonal projection function onto the space $\mathcal{X}$, Φ designs the standard Gaussian cumulative distribution function and $F_{\chi^2(k)}$ is the cumulative distribution function of a chi-squared variable with k degrees of freedom. By convention, for an intersection or an union from indices k to ℓ with $k > \ell$, the intersection or the union is over an empty set. In the same way, the set $\{k, \cdots, \ell\}$ is empty if $k > \ell$.

4.3 The main results.

In this section, we first present intuitions and key ideas that lead to the study of the FDR with respect to the penalty function calibration. Then, non-asymptotic bounds of the FDR are obtained in Theorem 4.1. Asymptotic behaviors of the FDR are then obtained when K tends to infinity. Finally, the particular case of the orthogonal design matrix is studied as an illustration of the main result.

4.3.1 Key ideas.

According to [Birgé and Massart, 2007], the penalty function (4.4) allows to achieve a non-asymptotic control of the PR if K is fixed strictly larger than 1. They discuss at length the best calibration of the hyperparameter K for a prediction point of view: the function $\mathrm{PR}(\hat{m}(K))$ with respect to $K > 0$ has huge values for $K < 1$, strongly decreases around 1, has low values for some $K > 1$, and then slowly increases in the over-fitting part. In practice, $K = 2$ is recommended since it allows the optimal asymptotic control of the PR. However, for a non-asymptotic point of view, other values of K close to 2 can give equivalent even better prediction performances. In this direction, we propose calibrating the hyperparameter K to maintain control of the PR while also ensuring control of the FDR. For this purpose, we propose to study simultaneously the functions $\mathrm{FDR}(\hat{m}(K))$ and $\mathrm{PR}(\hat{m}(K))$ for all $K > 0$ to find a value for K providing low values for both criteria simultaneously.

In this paragraph, we present a toy data set which is our common thread. We simulate $Y \sim \mathcal{N}(\beta^*, I_n)$ with $n = p = 50$, $D_{m^*} = 10$, $\beta^*_{10} = 2$ and $\forall j \in \{1, \cdots, 9\}$, $\beta^*_j \sim \mathrm{Unif}(\beta^*_{j+1} + 0.5, \beta^*_{j+1} + 1.5)$. Thus, coordinates of β^* are ranked in descending order and provide ordered active variables. Coordinate values of β^* are chosen to be higher than the noise and distant from each other. For the empirical estimations, we simulate 1000 data sets from this design. For the rest of the chapter, $\mathcal{D}$ denotes these 1000 data sets. We obtain the estimators $\hat{m}_d(K)$ on each dataset $d \in \mathcal{D}$ and since m^* is known, we evaluate $\mathrm{FDP}_d(\hat{m}_d(K))$ on each $d \in \mathcal{D}$ and for each $K > 0$. Then, the *empirical estimator of FDR*$(\hat{m}(K))$ is the average of the $\mathrm{FDP}_d(\hat{m}_d(K))$ across $\mathcal{D}$.
Concerning PR, 1000 new datasets, denoted by $\{d \in \tilde{\mathcal{D}}\}$, are generated using the same experimental design, as well as new $(\tilde{Y}_d)$ for $d \in \tilde{\mathcal{D}}$ generated from the model (4.1), and by using $(X_d)_{d \in \mathcal{D}}$ to respect the fixed design assumption. The selected models $\hat{m}_d(K)$ and the $\hat{\beta}_{\hat{m}(K),d}$

estimators are obtained from the training sets $\Big(Y_d, X_d\Big)_{d\in\mathcal{D}}$. The PR is evaluated from the validation sets $\Big(\tilde{Y}_d, X_d\Big)_{d\in\tilde{\mathcal{D}}}$ by the mean squared error: for each d and $\forall K > 0$,

$$\mathrm{MSE}_d(\hat{m}_d(K)) = \frac{1}{n}\sum_{i=1}^{n}\Big(\tilde{y}_{i_d} - \sum_{j=1}^{p} x_{ij_d}\hat{\beta}_{\hat{m}_d(K),j}\Big)^2. \tag{4.9}$$

The *empirical estimator of PR*$(\hat{m}(K))$ is the average of the $\mathrm{MSE}_d(\hat{m}_d(K))$ across $\mathcal{D}$.

In Figure 4.1, we plot the empirical estimators of $\mathrm{PR}(\hat{m}(K))$ and $\mathrm{FDR}(\hat{m}(K))$ on a regular grid of K. Firstly, to validate the quality of the empirical estimations, we compute the 95% asymptotic confidence interval obtained by the central limit theorem:

$$\big[\mathrm{FDR}(\hat{m}(K)) - 1.96\frac{\hat{\sigma}}{\sqrt{1000}}, \mathrm{FDR}(\hat{m}(K)) + 1.96\frac{\hat{\sigma}}{\sqrt{1000}}\big]$$

and

$$\big[\mathrm{PR}(\hat{m}(K)) - 1.96\frac{\hat{\sigma}}{\sqrt{1000}}, \mathrm{PR}(\hat{m}(K)) + 1.96\frac{\hat{\sigma}}{\sqrt{1000}}\big],$$

where $\hat{\sigma}$ is the unbiased empirical estimator of the standard deviation σ. Their width are tight (do not exceed 0.011 and 0.07 for respectively the FDR and the PR) meaning that the empirical estimations are closed to respectively $\mathrm{FDR}(\hat{m}(K))$ and $\mathrm{PR}(\hat{m}(K))$.
We illustrate our key ideas by using these two graphs. We observe that the $\mathrm{MSE}(\hat{m}(2))$ is small but the empirical $\mathrm{FDR}(\hat{m}(2))$ is large: the selected model contains several variables which are non active variables. Making the procedure more conservative would avoid their selection. For this purpose, one can increase the penalization (4.4) where the only free parameter is K. We observe in Figure 4.1 that for all $K \in [2, 3]$, the MSE values are kept low while the FDR value is lower for $K = 3$ than for $K = 2$. Increasing the constant K to limit the non active variable selection is motivated by the asymptotic point of view since $\hat{m}(2)$ leads to a zero value for the predictive risk asymptotically (AIC and the CP-Mallows penalties), while $\hat{m}(\log(n))$ leads to a consistency property for the set of active variables when n tends to infinity (BIC penalty). AIC and CP-Mallows give asymptotically the best set of variables for prediction performances, while BIC recovers the exact set of active variables asymptotically. According to [Yang, 2005], obtaining the asymptotic properties of the AIC and BIC penalties simultaneously is impossible, but this conclusion suggests that a value of K between 2 and $\log(n)$ can give reasonable (but not necessarily optimal) values for both PR and FDR when n and p are fixed (non-asymptotic framework). Hence, it seems possible to control the FDR value better while maintaining prediction performances similar to those from $K = 2$. We insist that we do not seek to simultaneously reach the non-asymptotic optimality of both criteria but only a good trade-off. In this way, we propose to study the function $\mathrm{FDR}\big(\hat{m}(K)\big)$ for all $K > 0$.

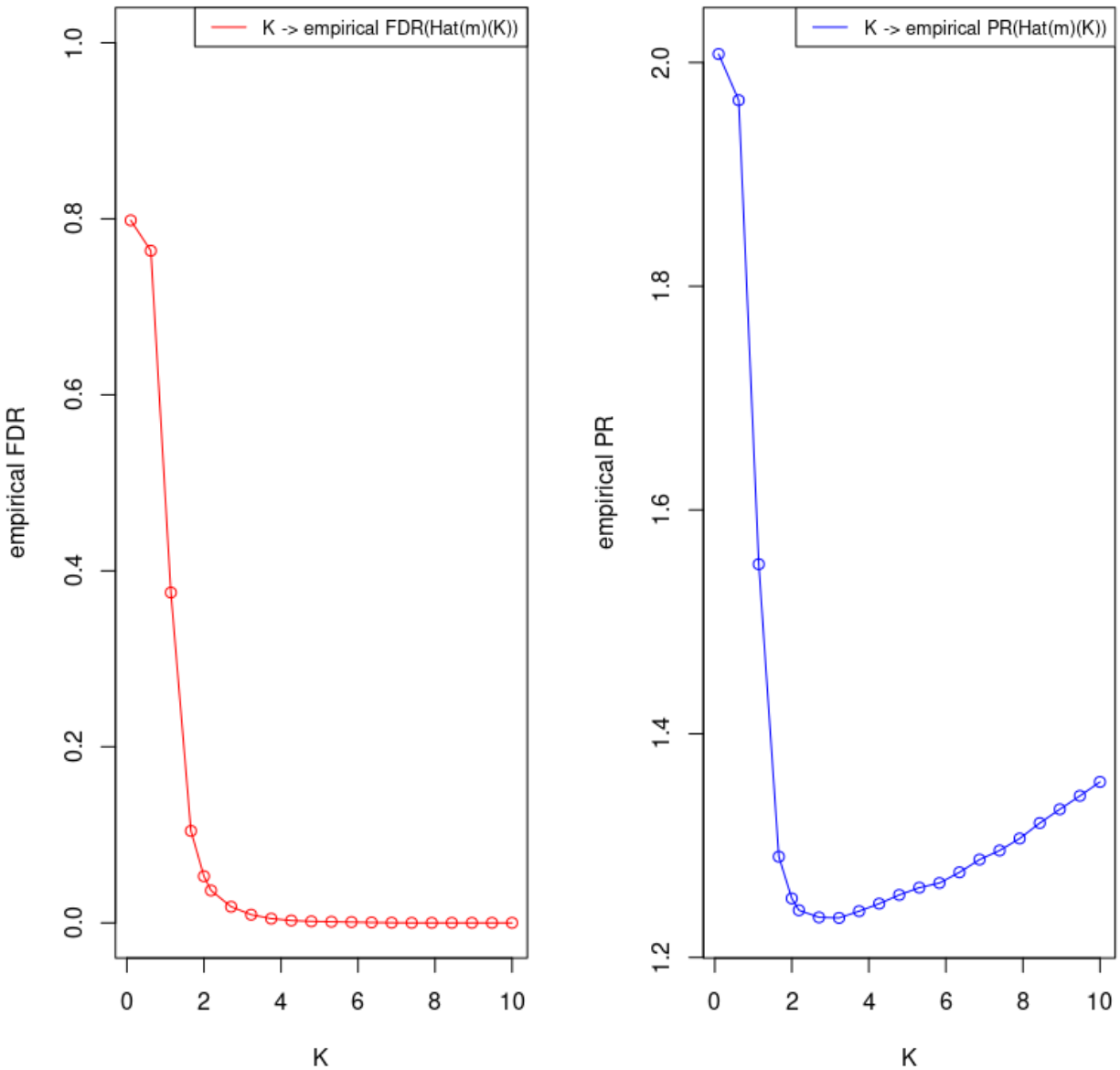


Figure 4.1: Curves of the empirical estimations of $\mathrm{FDR}\big(\hat{m}(K)\big)$ and $\mathrm{PR}\big(\hat{m}(K)\big)$ for all $K > 0$ by using 1000 datasets.

4.3.2 Bounds of the FDR in model selection.

This subsection presents all the theoretical results of this article. The variance σ^2 is supposed to be known.

4.3.2.1 FDR expression in model selection.

Let us consider the model collection (4.5) where $q = \min(n, p)$ and $m^* \in \mathcal{M}$. Let us assume that $\forall K > 0, \mathrm{crit}_K$ is injective on $\mathcal{M}$. Then, if $D_m^* = q$, $\mathrm{FDR}(\hat{m}(K)) = 0$, $\forall K > 0$. Otherwise, the $\mathrm{FDR}(\hat{m}(K))$ is expressed within the model selection procedure as: $\forall K > 0$,

$$\mathrm{FDR}(\hat{m}(K)) = \sum_{r=D_{m^*}+1}^{q} \frac{r - D_{m^*}}{r} \mathbb{P}\left(\left\{\bigcap_{\substack{\ell=0 \\ \ell \neq r}}^{q} \{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\}\right\}\right). \tag{4.10}$$

A detailed proof of (4.10) can be found in Subsection 4.5.1.
By using the decomposition $\left\{\bigcap_{\ell=0}^{r-1} \{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\}\right\} \bigcap \left\{\bigcap_{\ell=r+1}^{q} \{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\}\right\}$, we obtain the following proposition:

Proposition 4.1. *Let us consider the model collection (4.5) where $q = \min(n, p)$, $m^* \in \mathcal{M}$ and that $D_m^* < q$. Let us consider that $\forall K > 0, crit_K$ is injective on $\mathcal{M}$. Let us apply the Gram–Schmidt process to obtain $(u_1, \cdots, u_q)$ the orthonormal basis of $\mathbb{R}^q$ such that $Span(X_1, \cdots, X_j) = Span(u_1, \cdots, u_j)$, $\forall j \in \{1, \cdots, q\}$. If $q < n$, $(u_1, \cdots, u_q)$ is completed to an orthonormal basis on $\mathbb{R}^n$, noted $(u_1, \cdots, u_n)$.*
Then, $\forall K > 0$,

$$FDR(\hat{m}(K)) = \sum_{r=D_{m^*}+1}^{q} \frac{r - D_{m^*}}{r} \; P_r(K) \; Q_r(K, \beta^*, \sigma^2) \tag{4.11}$$

where for each $r \in \{D_{m^*} + 1, \cdots, q\}$,

$$P_r(K) = \mathbb{P}\left(\bigcap_{\ell=r+1}^{q} \left\{\sum_{k=r+1}^{\ell} Z_k^2 < K(\ell - r)\right\}\right), \tag{4.12}$$

$$\text{where} \quad Z_k \overset{i.i.d.}{\sim} \mathcal{N}(0, 1), \quad \forall k \in \{r + 1, \cdots, q\},$$

$$Q_r(K, \beta^*, \sigma^2) = \mathbb{P}\left(\bigcap_{\ell=0}^{r-1} \left\{\sum_{k=\ell+1}^{r} \langle Y, u_k \rangle^2 > K\sigma^2(r - \ell)\right\}\right). \tag{4.13}$$

Proof of Proposition 4.1 can be found in Subsection 4.5.1.

4.3.2.2 General bounds.

In (4.11), the $P_r(K)$ terms do not depend on the data. Conversely, the $Q_r(K, \beta^*, \sigma^2)$ terms depend on the data. Thus, to understand the behavior of the FDR function with respect to K, we propose to bound the $Q_r(K, \beta^*, \sigma^2)$ terms leading to the following theorem:

Theorem 4.1. *Let us consider the model collection (4.5) where $q = \min(n, p)$ and let us suppose that $m^* \in \mathcal{M}$ and $D_m^* < q$. The notation Φ stands for the standard gaussian cumulative distribution function and $F_{\chi^2(k)}$ is the cumulative distribution function of a chi-squared variable with k degrees of freedom. Let us assume that $\forall K > 0, crit_K$ is injective on $\mathcal{M}$. Let us apply the Gram–Schmidt process to obtain $(u_1, \cdots, u_q)$ the orthonormal basis of $\mathbb{R}^q$ such that $Span(X_1, \cdots, X_j) = Span(u_1, \cdots, u_j)$, $\forall j \in \{1, \cdots, q\}$. If $p < n$, $(u_1, \cdots, u_q)$ is naturally completed by the incomplete basis theorem in an orthonormal basis on $\mathbb{R}^n$, noted $(u_1, \cdots, u_n)$. Then, $\forall K > 0$, $\hat{m}(K)$ satisfies:*

$$b(K, \beta^*, \sigma^2) \leq FDR(\hat{m}(K)) \leq B(K, \beta^*, \sigma^2) \tag{4.14}$$

where $\Big[K \mapsto b(K, \beta^*, \sigma^2)\Big]$ and $\Big[K \mapsto B(K, \beta^*, \sigma^2)\Big]$ are two real-valued functions on $\mathbb{R}^+$: $\forall K > 0$,

$$b(K, \beta^*, \sigma^2) = \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} P_r(K) \ \underline{f}_r(K, \beta^*, \sigma^2) \right) \tag{4.15}$$

$$B(K, \beta^*, \sigma^2) = \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} P_r(K) \ \overline{f}_r(K, \beta^*, \sigma^2) \right) \tag{4.16}$$

with for all $K > 0$,

1. $$\text{for each } r \in \{D_{m^*} + 1, \cdots, q\}, \ P_r(K) = \mathbb{P}\Bigg(\bigcap_{\ell=r+1}^{q} \Big\{ \sum_{k=r+1}^{\ell} Z_k^2 < K(\ell - r) \Big\} \Bigg),$$
$$\text{where } \ Z_k \overset{i.i.d.}{\sim} \mathcal{N}(0, 1), \quad \forall k \in \{r + 1, \cdots, q\}.$$

2. $\Big(\underline{f}_r\Big)_{r \in \{1, \cdots, q\}}$ *and* $\Big(\overline{f}_r\Big)_{r \in \{D_{m^*}+1, \cdots, q\}}$ *are real-valued functions on* $\mathbb{R}^+$ *defined by:*

$$\forall r \in \{D_{m^*}+1, \cdots, q\},\ \forall \ell \in \{2, \cdots, q\},$$

$$\overline{f}_r(K, \beta^*, \sigma^2)) = 1 - \max\Bigg(\max_{\ell \in \{1, \cdots, r-D_{m^*}\}} \Big(F_{\chi^2(\ell)}(\ell K) \Big),$$

$$\max_{\ell \in \{r-D_{m^*}+1, \cdots, r\}} \Bigg(F_{\chi^2(\ell)} \Big(\frac{\ell K}{2} - \sum_{k=r-\ell+1}^{D_{m^*}} \frac{\langle X\beta^*, u_k \rangle^2}{\sigma^2} \Big) \Big) \Bigg)$$

$$\underline{f}_1(K, \beta^*) = G_1$$

$$\underline{f}_\ell(K, \beta^*, \sigma^2) = G_\ell + H_\ell\ \underline{f}_{\ell-1}(K, \beta^*)$$

$$\textit{with for}\ \ \ell \in \{D_{m^*}+1, \cdots, r\}:\ \ G_\ell = 2\Big(1 - \Phi\big(\sqrt{\ell K}\big)\Big)$$

$$H_\ell = 2\Big(\Phi\big(\sqrt{\ell K}\big) - \Phi\big(\sqrt{K}\big)\Big)$$

$$\textit{for}\ \ \ell \in \{1, \cdots, D_{m^*}\}:\ \ G_\ell = 2 - \Bigg(\Phi\Big(\sqrt{\ell K} - \frac{\langle X\beta^*, u_\ell \rangle}{\sigma}\Big) + \Phi\Big(\sqrt{\ell K} + \frac{\langle X\beta^*, u_\ell \rangle}{\sigma}\Big)\Bigg)$$

$$\textit{for}\ \ \ell \in \{2, \cdots, D_{m^*}\}:\ \ H_\ell = \Phi\Big(\sqrt{\ell K} - \frac{\langle X\beta^*, u_\ell \rangle}{\sigma}\Big) + \Phi\Big(\sqrt{\ell K} + \frac{\langle X\beta^*, u_\ell \rangle}{\sigma}\Big) -$$

$$\Bigg(\Phi\Big(\sqrt{K} - \frac{\langle X\beta^*, u_\ell \rangle}{\sigma}\Big) + \Phi\Big(\sqrt{K} + \frac{\langle X\beta^*, u_\ell \rangle}{\sigma}\Big)\Bigg).$$

Proof of Theorem 4.1 can be found in Subsection 4.5.2.
Hence, although the model selection procedure is built for prediction performances, it is also possible to derive some results on the FDR thanks to these bounds. The quantities $P_r(K)$ can be calculated once and for all without any dataset. The other terms $\underline{f}_r(K, \beta^*, \sigma^2))$ and $\overline{f}_r(K, \beta^*, \sigma^2))$ only involve some evaluations of cumulative distribution functions of the standard Gaussian and chi-squared variables; they have a fully explicit form; and they depend only on the model parameters. Note that some of them depend on the signal-to-noise ratio, as usual in statistics.

4.3.2.3 In a no noise framework, FDR is strictly positive.

The following corollary gives a lower bound of the FDR which does not depend on σ^2.

Corollary 4.1. *Under the assumptions and definitions of Theorem 4.1, $\forall K > 0$:*

$$FDR(\hat{m}(K)) \geq \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} P_r(K)\ \frac{2\sqrt{2}}{\sqrt{\pi}\Big(\sqrt{rK} + \sqrt{rK+4}\Big)} e^{-\frac{rK}{2}} \right) > 0.$$

Proof of Corollary 4.1 can be found in Subsection 4.5.3.
Hence, $\mathrm{FDR}(\hat{m}(K)) > 0$ for all $K > 0$ and whatever $\sigma > 0$. This is not surprising since the $\mathrm{FDR}(\hat{m}(K))$ is strictly positive even in the simplest case of no noise level. Indeed, when $\sigma = 0$, $Y = X\beta^*$ and the minimization in (4.7) is reduced to the least-squares minimization. So, $\hat{\beta}_{m^*} = \beta^*$ and the associated least-squares criterion is null. However, for $m \in \mathcal{M}$ such that $m^* \subset m$, $\hat{\beta}_m$ can also cancel the least-squares criterion. The selected model $\hat{m}$ can then be strictly larger than m^* and so contains non active variables. So, the $\mathrm{FDP}(\hat{m})$ is strictly positive, and by taking the expectation, $\mathrm{FDR}(\hat{m}) > 0$. The larger σ, the larger the FDR, so for $\sigma > 0$, the event $m^* \subset \hat{m}$ happens with a non-zero probability providing a strictly positive FDR as well.

4.3.2.4 Asymptotic analysis.

The following corollary gives the asymptotic behavior of the FDR function when K tends to infinity.

Corollary 4.2. *Under the assumptions and definitions of Theorem 4.1, the $FDR(\hat{m}(K))$ function tends to 0 when K tends to infinity and verifies $\forall \eta > 0$,*

$$FDR(\hat{m}(K)) = \underset{K \longrightarrow +\infty}{o}\left(e^{-K(\frac{1}{2}-\eta)}\right). \tag{4.17}$$

Furthermore, $\forall \eta > 0,\ \exists C_\eta > 0,\ \exists L_\eta > 0,\ \forall K > L_\eta$, we have:

$$FDR(\hat{m}(K)) \geq C_\eta e^{-K\left(\frac{D_{m^*}+1+2\eta}{2}\right)}. \tag{4.18}$$

So, $\forall \varepsilon > 0$,

$$\begin{aligned} -\frac{D_{m^*}}{2} - \frac{1}{2} - \varepsilon \ \leq \ & \liminf_{K \longrightarrow +\infty} \frac{1}{K} \log\left(FDR(\hat{m}(K))\right) \\ & \limsup_{K \longrightarrow +\infty} \frac{1}{K} \log\left(FDR(\hat{m}(K))\right) \ \leq \ -\frac{1}{2} + \varepsilon. \end{aligned} \tag{4.19}$$

Proof of Corollary 4.2 can be found in Subsection 4.5.4.

Remark 4.1. With no signal ($\beta^* = 0$ and so $D_{m^*} = 0$), the asymptotic bounds in (4.19) are respectively $-\frac{1}{2} - \varepsilon$ and $-\frac{1}{2} + \varepsilon$ leading to a behavior of the $\mathrm{FDR}(\hat{m}(K))$ like $e^{-\frac{K}{2}}$ for K large enough.

Remark 4.2. The asymptotic upper and lower bounds (4.17) and (4.18) are true whatever the value of $\sigma^2 > 0$. It is possible to obtain the following sharpest asymptotic upper bound: $\forall \tilde{\eta} > 0$,

$$\mathrm{FDR}(\hat{m}(K)) = o\left(e^{-\left(K\frac{(D_{m^*}+1-\tilde{\eta})}{4} - \frac{1}{2\sigma^2}\sum_{k=1}^{D_{m^*}}\langle X\beta^*, u_k\rangle^2\right)}\right) \tag{4.20}$$

in the asymptotic regime where $K \longrightarrow +\infty$ and $\sigma \longrightarrow 0$ with $\frac{1}{\sigma} = \underset{\sigma\longrightarrow 0}{o}(\sqrt{K})$. This regime means that the noise amplitude σ can tend to 0 but the convergence rate has to be slower than the one of K. The reader can find the proof is in Section 4.5.

The $\mathrm{FDR}(\hat{m}(K))$ tends to 0 when K tends to $+\infty$ and Corollary 4.2 shows that the convergence is at an exponential rate. Equation (4.18) suggests that the exponential rate is the optimal convergence rate order. Moreover, although this result is asymptotic, (4.17) suggests that since the convergence rate is exponential, there is no need to go far from $K = 2$ to have a reasonably small FDR.

4.3.3 Illustrations of Theorem 4.1 in the orthogonal case of Corollary 4.3.

We propose to analyze the particular case of the orthogonal design matrix since it leads to simplified forms for the FDR bounds easy to implement.

Corollary 4.3 (Application on the orthogonal case)**.** *Under assumptions of Theorem 4.1 and by assuming that $(X_1, \cdots, X_q)$ are orthonormal for $\langle ., . \rangle$, then:*
$\forall K > 0, \hat{m}(K)$ satisfies the same inequalities as (4.14) with:
for $\ell \in \{1, \cdots, D_{m^*}\}$:

$$G_\ell = 2 - \left(\Phi\left(\sqrt{\ell K} - \frac{\beta^*_\ell}{\sigma}\right) + \Phi\left(\sqrt{\ell K} + \frac{\beta^*_\ell}{\sigma}\right)\right),$$

for $\ell \in \{2, \cdots, D_{m^*}\}$:

$$H_\ell = \Phi\left(\sqrt{\ell K} - \frac{\beta^*_\ell}{\sigma}\right) + \Phi\left(\sqrt{\ell K} + \frac{\beta^*_\ell}{\sigma}\right) - \left(\Phi\left(\sqrt{K} - \frac{\beta^*_\ell}{\sigma}\right) + \Phi\left(\sqrt{K} + \frac{\beta^*_\ell}{\sigma}\right)\right)$$

and for all $r \in \{D_{m^*} + 1, \cdots, q\}$:

$$\overline{f}_r(K, \beta^*, \sigma^2)) = 1 - \max\left(\max_{\ell\in\{1,\cdots,r-D_{m^*}\}}\left(F_{\chi^2(\ell)}(\ell K)\right), \max_{\ell\in\{r-D_{m^*}+1,\cdots,r\}}\left(F_{\chi^2(\ell)}\left(\frac{\ell K}{2} - \sum_{k=r-\ell+1}^{D_{m^*}}\frac{\beta^{*2}_k}{\sigma^2}\right)\right)\right).$$

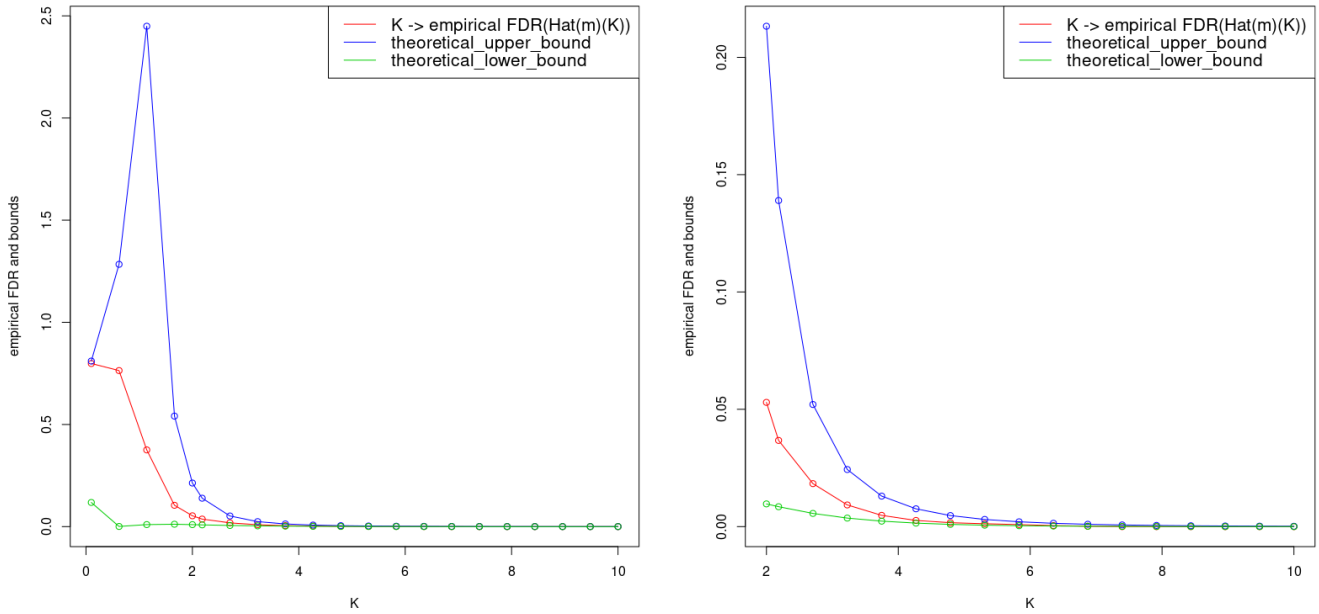


Figure 4.2: Curves of the empirical estimation of the FDR and the terms $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$ under the orthogonal design of Corollary 4.3. Right: curves are plotted only for $K \geq 2$.

Proof of Corollary 4.3 can be found in Subsection 4.5.5.

In Figure 4.2, we plot the empirical estimation of the FDR$(\hat{m}(K))$ and the functions $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$ on a grid of $K > 0$ and under the orthogonal design of Corollary 4.3, by using the toy data sets $\mathcal{D}$ described in Subsection 4.3.1. As expected, the empirical FDR$(\hat{m}(K))$ approaches 0 when K increases and the convergence rate seems to be exponential, which is consistent with Corollary 4.2. Moreover, the curves of $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$ frame the empirical FDR. If we focus on what happens when $K \geq 2$, their decrease is also exponential and the difference between the three curves becomes quickly negligible. In Appendix 4.7 and in supplementary material available in [1], we propose a more detailed study of the terms $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$. More precisely, we evaluate the inequalities established in Proposition 4.1 until Theorem 4.1 through a comparison of $\underline{f}_r$ and $\overline{f}_r$ with respect to the quantity Q_r. We observe curves almost overlap, suggesting that the used inequalities are not too brutal. In Appendix 4.7, curves are plotted for $\sigma^2 = 1$. In supplementary material available in [2], the reader can find curves for $\sigma^2 = 0.1$ and $\sigma^2 = 4$.

[1] `https://sites.google.com/view/placroix/research`
[2] `https://sites.google.com/view/placroix/research`

4.4 Trade-off between the PR and the FDR controls.

While bounds $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$ are fully implementable, they depend on β^* and σ^2, unknown in practice. In this section, we propose to consider the constant K as a hyperparameter to calibrate in the model selection procedure (4.7). Our strategy consists in studying simultaneously the curves of the functions $\mathrm{FDR}\big(\hat{m}(K)\big)$ and $\mathrm{PR}\big(\hat{m}(K)\big)$ for all $K > 0$ to calibrate the hyperparameter K for keeping a low value of the PR while gaining a low value of the FDR to guarantee respectively prediction performances and to avoid non active variable in the selected variables subset. Firstly, theoretical terms are estimated from the data (Subsection 4.4.1) to secondly calibrate the hyperparameter K (Subsection 4.4.2).

4.4.1 Data-driven estimation of the theoretical terms.

An empirical approach can reasonably estimate the PR and FDR functions when a lot of data sets are available. When only one dataset is available, an alternative is to replace the theoretical terms by data-driven ones.

4.4.1.1 The predictive risk.

Instead of evaluating the mean squared error (4.9) requiring separating the data into a training set and a validation set, we propose to use the entire available dataset to both apply the model selection procedure and evaluate the theoretical predictive risk. Let us observe that for all $K > 0$ and $K' > 0$:

$$\begin{aligned}
&\mathbb{E}[||Y - X\hat{\beta}_{\hat{m}(K)}||_2^2] - \mathbb{E}[||Y - X\hat{\beta}_{\hat{m}(K')}||_2^2] \\
&= \mathbb{E}[||X\hat{\beta}_{\hat{m}(K)}||_2^2] - \mathbb{E}[||X\hat{\beta}_{\hat{m}(K')}||_2^2] + 2\mathbb{E}[\langle Y, X\hat{\beta}_{\hat{m}(K')} - X\hat{\beta}_{\hat{m}(K)}\rangle] \\
&= \mathbb{E}[||X\hat{\beta}_{\hat{m}(2)} - X\hat{\beta}_{\hat{m}(K)}||_2^2] - \mathbb{E}[||X\hat{\beta}_{\hat{m}(2)} - X\hat{\beta}_{\hat{m}(K')}||_2^2] \\
&\quad + 2\mathbb{E}[\langle X\hat{\beta}_{\hat{m}(K)} - X\hat{\beta}_{\hat{m}(K')}, X\hat{\beta}_{\hat{m}(2)}\rangle] + 2\mathbb{E}[\langle Y, X\hat{\beta}_{\hat{m}(K')} - X\hat{\beta}_{\hat{m}(K)}\rangle] \\
&= \mathbb{E}[||X\hat{\beta}_{\hat{m}(2)} - X\hat{\beta}_{\hat{m}(K)}||_2^2] - \mathbb{E}[||X\hat{\beta}_{\hat{m}(2)} - X\hat{\beta}_{\hat{m}(K')}||_2^2] \\
&\qquad - 2\mathbb{E}[\langle X\hat{\beta}_{\hat{m}(2)} - Y, X\hat{\beta}_{\hat{m}(K')} - X\hat{\beta}_{\hat{m}(K)}\rangle].
\end{aligned} \tag{4.21}$$

The constant 2 allows to get the optimal asymptotic control of (4.8). Consequently, firstly, $||Y - X\hat{\beta}_{\hat{m}(2)}||_2$ is close to 0. Secondly, $X\hat{\beta}_{\hat{m}(2)}$ is close to $\Pi_{\mathrm{Im}(X)}(Y)$ and so $X\hat{\beta}_{\hat{m}(2)} - Y$ almost belongs to the subspace $\mathrm{Im}(X)^{\perp}$, unlike $X\hat{\beta}_{\hat{m}(K)}$ and $X\hat{\beta}_{\hat{m}(K')}$ belonging to $\mathrm{Im}(X)$. Hence, in (4.21), the term $\mathbb{E}[\langle X\hat{\beta}_{\hat{m}(2)} - Y, X\hat{\beta}_{\hat{m}(K')} - X\hat{\beta}_{\hat{m}(K)}\rangle]$ is close to 0 and is negligible compared to the two others.
Equalities in (4.21) show that the dynamics of $\mathbb{E}[||X\hat{\beta}_{\hat{m}(2)} - X\hat{\beta}_{\hat{m}(K)}||_2^2]$ with respect to $K > 0$ is close to the one of $\mathbb{E}[||Y - X\hat{\beta}_{\hat{m}(K)}||_2^2]$. So, the constant K minimizing $\mathbb{E}[||X\hat{\beta}_{\hat{m}(2)} - X\hat{\beta}_{\hat{m}(K)}||_2^2]$ and the one minimizing $\mathbb{E}[||Y - X\hat{\beta}_{\hat{m}(K)}||_2^2]$ are almost equal.

Finally, to evaluate the prediction performances, we use the following term that we call estimated risk for the sequel:

$$\widehat{\mathrm{PR}}(\hat{m}(K)) = \frac{1}{n}\sum_{i=1}^{n}\Big(\big(X\hat{\beta}_{\hat{m}(2)}\big)_i - \big(X\hat{\beta}_{\hat{m}(K)}\big)_i\Big)^2. \tag{4.22}$$

4.4.1.2 The bounds of the FDR.

Theorem 4.1 gives a lower bound and an upper bound of $\mathrm{FDR}\big(\hat{m}(K)\big)$ for all $K > 0$. These bounds are explicit and easily implementable.
The P_r quantities do not depend on the data as soon as r is given. For each $1 \leq r \leq q$, P_r is calculated by generating 5000 independent standard Gaussian vectors $\Big(Z_k\Big)_{k\in\{r+1,\cdots,q\}}$ and by counting for each vector the number of times that $Z_k^2 < K(\ell - r)$ for each $\ell \in \{r+1, \cdots, q\}$. For all $K > 0$, the $\underline{f}_r(K, \beta^*, \sigma^2)$ and $\overline{f}_r(K, \beta^*, \sigma^2)$ quantities depend on β^* and σ^2, both unknown.

The slope heuristic to estimate σ^2. Slope heuristics have been introduced in [Birgé and Massart, 2007] to overcome the problem of σ^2 unknown in penalty functions. By using the definition of the empirical contrast γ_n defined by: $\forall \beta \in \mathbb{R}^p$,

$$\gamma_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2,$$

the principle is that when D_m is large enough, $\gamma_n(\hat{\beta}_m)$ is equal to $-\frac{1}{2n}K\sigma^2 D_m$ plus an additive constant independent of n and m. Hence, $\hat{\sigma}^2$ is calibrated from the dataset by estimating the multiplicative coefficient of the affine behavior between $-\gamma_n(\hat{\beta}_m)$ and $-\frac{K}{2n}D_m$ for D_m large enough. We use the function capushe of the R package capushe (version 1.1.1) [Baudry et al., 2012] to apply the slope heuristic. The parameters are set to the default values. Figure 4.3 displays the histogram of the $\hat{\sigma}^2$ values obtained by the slope heuristics applied on 100 sets of $\mathcal{D}$. The median equals 0.96 with a mean and a standard deviation equals 1 and 0.45. So, the values of $\hat{\sigma}^2$ are relatively closed to the true variance σ^2 equals 1. This validates the use of the slope heuristics to estimate σ^2.

Some substitutes of β^*. Concerning β^*, we propose to replace it with an estimator of β^*. According to [Birgé and Massart, 2007], $\hat{\beta}_{\hat{m}(K)}$ is a good estimator of β^* in a predictive point of view when K is equal or close to 2. So, an idea is to evaluate the $\underline{f}_r$ and $\overline{f}_r$ functions by using such an estimator. In this article, we propose to evaluate the lower and upper bounds $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$ of Theorem 4.1 by using $\hat{\beta}_{\hat{m}(\tilde{K})}$ for $\tilde{K} \in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, \log(n)\}$. The first extreme value 1 stands for the critic constant to get the non-asymptotic control of the PR [Birgé and Massart, 2001]. The second extreme value $\log(n)$ stands for the constant appearing in the BIC penalty to get consistency for the selected variables set [Schwarz et al.,

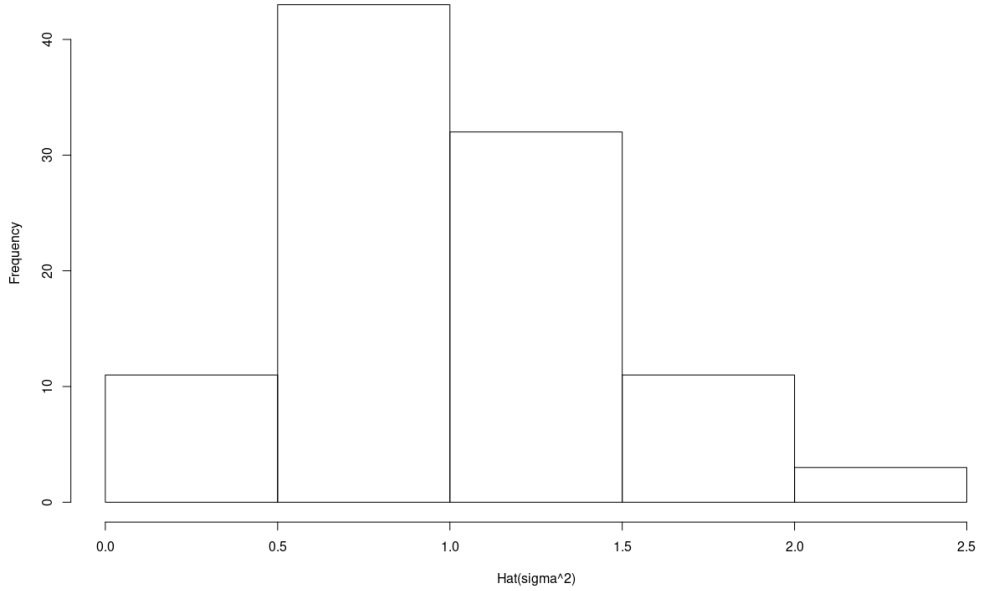


Figure 4.3: Histogram over 100 data sets of the values of $\hat{\sigma}^2$ obtained by the slope heuristic with the function capushe of the R package capushe (version 1.1.1)

1978]. As for $K = 2$, it corresponds to the AIC penalty, which leads to optimal asymptotic control of the PR [Akaike, 1973] and is the constant usually used in practice. The other constants close to 2 are arbitrarily fixed.
We denote $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ the estimations of $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$. As these bounds $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are random, to relevantly choose the hyperparameter $\tilde{K}$, we propose to evaluate these replacements on a large experimental study under the orthogonal case of Corollary 4.3. This large experimental study uses $\mathcal{D}$ but also other datasets with the same experimental design as Subsection 4.3.1 but with different values for D_{m^*}, β^*, n and σ^2. Table 4.1 describes the datasets.

The scenario (i) tests the impact of the degree of sparsity. The last two parameters β^* in scenario (ii) have close non-zero coefficients, making the construction of the variables order more complex than the first β^*. Moreover, the second β^* in (ii) has some of its non-zero coefficients smaller than the amplitude of the noise σ, which complicates the distinction between non-zero coefficients and the others. The scenario (iii) studies the consequences of the high-dimension since the number of available observations is smaller, equal or larger than $p = 50$. Note that for a fair comparison, the 30-observations are included in the 50-observations, included itself in the 300-observations and the seed of the random number generator is identically fixed between each scenario. The last scenario (iv) tests different values of σ^2 for the noise amplitude impact.

For the rest of this subsection, we propose a complete analysis of bounds $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and

Scenario with $p = 50$	Active variables number	Non-zero coefficients amplitude in β^*	Observations number	Noise amplitude
(i) Sparsity	$D_{m^*} \in \{1, 10, 20\}$	$\beta^*_{D_{m^*}} = 2$, $\forall j \in \{1, \cdots, D_{m^*} - 1\}$ $\beta^*_j \sim \text{Unif}(\beta^*_{j+1} + 0.5, \beta^*_{j+1} + 1.5)$	$n = 50$	$\sigma^2 = 1$
(ii) Complexity	$D_{m^*} = 10$	$\beta^*_{10} = 2$ with $\forall j \in \{1, \cdots, 9\}$ $\beta^*_j \sim \text{Unif}(\beta^*_{j+1} + 0.5, \beta^*_{j+1} + 1.5)$ or $\beta^*_{10} = \frac{2}{10}$ with $\forall j \in \{1, \cdots, 9\}$, $\beta^*_j \sim \text{Unif}(\beta^*_{j+1} + 0.05, \beta^*_{j+1} + 0.15)$ or $\beta^*_{10} = 2$ with $\forall j \in \{1, \cdots, 9\}$ $\beta^*_j \sim \text{Unif}(\beta^*_{j+1} + 0.05, \beta^*_{j+1} + 0.15)$	$n = 50$	$\sigma^2 = 1$
(iii) High-dimension	$D_{m^*} = 10$	$\beta^*_{D_{m^*}} = 2$, $\forall j \in \{1, \cdots, D_{m^*} - 1\}$ $\beta^*_j \sim \text{Unif}(\beta^*_{j+1} + 0.5, \beta^*_{j+1} + 1.5)$	$n \in \{30, 50, 300\}$	$\sigma^2 = 1$
(iv) Noise	$D_{m^*} = 10$	$\beta^*_{D_{m^*}} = 2$, $\forall j \in \{1, \cdots, D_{m^*} - 1\}$ $\beta^*_j \sim \text{Unif}(\beta^*_{j+1} + 0.5, \beta^*_{j+1} + 1.5)$	$n = 50$	$\sigma^2 \in \{0.1, 1, 4\}$

Table 4.1: Description of scenarios for the four experimental data sets generated identically to those of the toy data set described in Subsection 4.3.1 but with some other parameters.

$\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ for $\tilde{K} \in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, \log(n)\}$ on these four scenarios. However, for the sake of conciseness, only curves from the toy data set $\mathcal{D}$ described in Subsection 4.3.1 are plotted at the heart of the Chapter; only curves from scenarios (i) and (ii) are plotted in Appendix 4.8; and all others are provided in the supplementary material available in [3].
To obtain $\hat{\beta}_{\hat{m}(\tilde{K})}$, the model collection is built on one dataset generated independently of $\mathcal{D}$ and respecting again the experimental design of Section 4.3.1. Then, the problem (4.7) is solved on this dataset and $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are evaluated.
In Figures 4.4, 4.5 (for the toy data set) and 4.20, 4.27 (for data from scenarios (i) and (ii)), the empirical estimation of the $\text{FDR}\big(\hat{m}(K)\big)$ on a grid of $K > 0$, as well as quantities $b(K, \beta^*, \sigma^2)$, $B(K, \beta^*, \sigma^2)$, $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are plotted.
To evaluate the error committed by replacing $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$ with their estimation

[3] `https://sites.google.com/view/placroix/research`

$\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, we propose the relative changes: $\forall K > 0$,

$$\frac{\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2) - b(K, \beta^*, \sigma^2)}{b(K, \beta^*, \sigma^2)}$$

for the lower bound and:

$$\frac{\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2) - B(K, \beta^*, \sigma^2)}{B(K, \beta^*, \sigma^2)}$$

for the upper bound. We expect to have the relative change as close to 0 as possible, positive values for the upper bounds to ensure that $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ is above $B(K, \beta^*, \sigma^2)$ and so above the FDR, and negative values for the lower bounds to ensure that $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ is below $B(K, \beta^*, \sigma^2)$ and so below the FDR.
To take into account of the randomness of the $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ terms, we generate 100 data sets, that we denote $\tilde{\mathcal{D}}$, independently of $\mathcal{D}$ and respecting again the experimental design of Section 4.3.1. With each one, $\hat{\beta}_{\hat{m}(\tilde{K})}$ is calculated by solving (4.7) given an evaluated bound function with respect to K. Then, the variance of the 100 bounds $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ is calculated and the relative standard deviation (standard deviation divided by the mean) is obtained for all K. We expect the relative standard deviation to be as close to 0 as possible. Relative changes and relative standard deviations for the $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ bounds are plotted in Figures 4.6,- 4.9 for the toy sets, in Figures 4.14- 4.19 and 4.21- 4.26 for scenarios (i) and (ii), in the supplementary material available in [4] for scenarios (iii) and (iv).

The lower bounds: Concerning the lower bounds $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, excepted for $\tilde{K} = 1$, the relative change values are positive until achieving more than 200% for large K and with the toy data set (Figure 4.6). It is undesirable since this means that estimated lower bounds curves can be larger than the theoretical one. Moreover, in some non-negligible times, the estimated lower bounds can be larger than the empirical FDR estimation one, as soon as $K \geq 2$. The relative standard deviation functions increase quickly whatever the value of $\tilde{K}$ and for all scenarios. For $\tilde{K} = 1$, the relative standard deviation function increases quickly until exceeding almost 2 for $K = 2$, almost 3 for $K = 5$ and almost 4 for $K = 8$. It suggests that fluctuations around the mean are not negligible (Figure 4.7).

The upper bounds: Concerning the upper bounds $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, excepted for $\tilde{K} = 1$, the relative change functions are always positive with the toy data set (Figure 4.8), so the $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ curves are above the $B(K, \beta^*, \sigma^2)$ ones and so also above the empirical FDR one. Moreover, excepted for $\tilde{K} = 1$, the relative change values don't exceed 0.11% meaning that the $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ curves are closed to $B(K, \beta^*, \sigma^2)$ for all $K > 0$. Concerning the others scenarios (see Figures 4.14, 4.15, 4.16, 4.21, 4.22 and 4.23 and the supplementary material available in [5]), values are small but can be negative, meaning that the theoretical upper bound exceeds the

[4] `https://sites.google.com/view/placroix/research`
[5] `https://sites.google.com/view/placroix/research`

estimated ones. This is undesirable, but it happens very rarely for $\tilde{K} \geq 4$ (and more and more frequently when $\tilde{K}$ decreases). In this rare situation for $\tilde{K} \geq 4$, values are low enough (smaller than -20%) to ensure that the empirical FDR estimation curves (whose confidence intervals are really tight with a width no more than 0.02) do not exceed the $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ terms. Concerning the relative standard deviation functions (see Figures 4.9, 4.17, 4.18, 4.19, 4.24, 4.25 and 4.26 and the supplementary material available in [6]), the larger $\tilde{K}$, the smaller the values, except for the scenario (ii) with the third β^* configuration: values increase after $\tilde{K} \geq 4.5$ (Figure 4.26). For $\tilde{K} \geq 3.5$, the relative standard deviation values are around 0.2 for all the scenarios except for scenario (ii) with the second β^* (can achieve 0.8, see Figure 4.24) and with the third β^* (can achieve 1, see Figure 4.26). Thus, for a value of $\tilde{K} \in \{3.5, \log(n), 4, 4.5, 5\}$ and eventually, except for the two extreme scenarios, fluctuations around the mean are small, meaning that the upper bound estimations are stable. Comparisons of impacts of parameter variation in results are proposed in Appendix 4.8.

To conclude, we drop the lower bound to implement our data-driven hyperparameter calibration since we can not guarantee that the studied $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions are smaller than the theoretical FDR one. It is not a problem since an upper bound is sufficient to get control of the FDR.
Concerning the upper bounds, the best results are obtained with the hyperparameter $\tilde{K} = 4$, where the relative change values are almost always positive. Moreover, values are always small enough to guarantee that the $\tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ are larger than the theoretical FDR and the relative standard deviation values are the smallest ones whatever the scenarios.
The value of the hyperparameter $\tilde{K} = 4$ is not surprising since to get an $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ larger than the theoretical upper one, the value of $D_{\hat{m}}$ has to be small enough in (4.16) to have more terms in the sum. So, the penalization function has to be large enough in (4.7). We have no hope to get better results for $K > 5$ since there is no difference, or poorer result, with $\tilde{K} = \{4.5, 5\}$.

Hence, to implement the data-driven hyperparameter calibration, we propose to replace β^* by $\hat{\beta}_{\hat{m}(4)}$.

4.4.2 A completely data-dependent calibration of the hyperparameter K in model selection procedure.

We propose a data-driven calibration of K by studying the estimated risk (4.22) and the $\tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions. The goal is to get a value of K guaranteeing a low value of both theoretical PR and FDR.

The data-driven hyperparameter calibration we propose is summarized in algorithm 1:

[6] `https://sites.google.com/view/placroix/research`

Algorithm 1: Algorithm to calibrate K

1. Choose α the threshold for the FDR control.
2. Compute $I_1 = \Big\{K \geq 2,\ \tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)\ \in]0, \alpha[\Big\}$.
3. Compute $I_2 = \Big\{K \geq 2,\ \widehat{\mathrm{PR}}(\hat{m}(K)) \approx \widehat{\mathrm{PR}}(\hat{m}(2))\Big\}$.
4. If $I_1 \cap I_2 \neq \emptyset$, choose $\min\Big\{K,\ K \in I_1 \cap I_2\Big\}$. ;

   Otherwise, choose $\min\Big\{K, K\ \in I_1\Big\}$ or take a larger value of α.

This algorithm offers control of the PR under a constraint on the FDR (smaller than a given threshold α). The interval I_1 is chosen to avoid a null-value for FDR, which could lead to an empty selected set, which is undesirable in practice.

We propose to detail the application of this algorithm in the toy set. In this example, we choose $\alpha = 0.05$. On Figure 4.10 are plotted the empirical estimation functions of $\mathrm{PR}\big(\hat{m}(K)\big)$ and $\mathrm{FDR}\big(\hat{m}(K)\big)$ for all $K > 0$ and from $\mathcal{D}$ as described in Subsection 4.3.1, with the plots of the estimated risk (4.22) and the $\tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions with the toy data set. Since the 95% asymptotic confidence intervals are very tight (don't exceed 0.011 and 0.07 for respectively the FDR and the PR), we can rely on curves of the FDR and PR empirical estimation as references to validate our data-driven calibration. Concerning the estimated risk (4.22) and the $\tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions, they are the only ones available in practice.
For K between 3.3 and 10, the values of the $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are non zero, lower than 0.05 and decrease with K. So, $I_1 = [3.3, 10]$. The estimated risk is minimal for $K \in [1.2, 5.8]$. So $I_2 = [1.2, 5.8]$. The intersection of I_1 and I_2 is $[3.3, 5.8]$. Our proposed algorithm leads to $K = 3.3$, the minimum of $I_1 \cap I_2$.
When we focus on the reference curves, we observe that the smallest values of the PR are obtained for K between 2 and 4.2 and the non-zero values smaller than 0.05 for the FDR are obtained for K between 2.2 and 6.8. Hence, a value of K in the interval $[2.2, 4.2]$ controls both the theoretical FDR and PR quantities.
The obtained constant $K = 3.3$ belongs to $[2.2, 4]$, the interval for low values of both theoretical FDR and PR. Hence, this data-driven hyperparameter calibration allows adding a FDR control in the model selection procedure while conserving a PR control.
With the other scenarios (see Figures 4.20, 4.27 and the supplementary material available in [7]) our data-driven calibration gives $K = 2.7$ for scenario (i) with $D_m^* = 20$; gives $K = 4.8$ for scenario (i) with $D_m^* = 1$ and for scenario (ii) with the second β^* configuration; and gives $K = 3.3$ for all the others. Thus, the calibrated hyperparameter K in all scenarios is strictly larger than the usual constant 2. The phenomenon that the value of the calibrated K varies with D_m^* is consistent with the fact that the larger the D_{m^*}, the larger the dimension of the

[7] `https://sites.google.com/view/placroix/research`

selected model should be and therefore, the smaller the penalty should be.

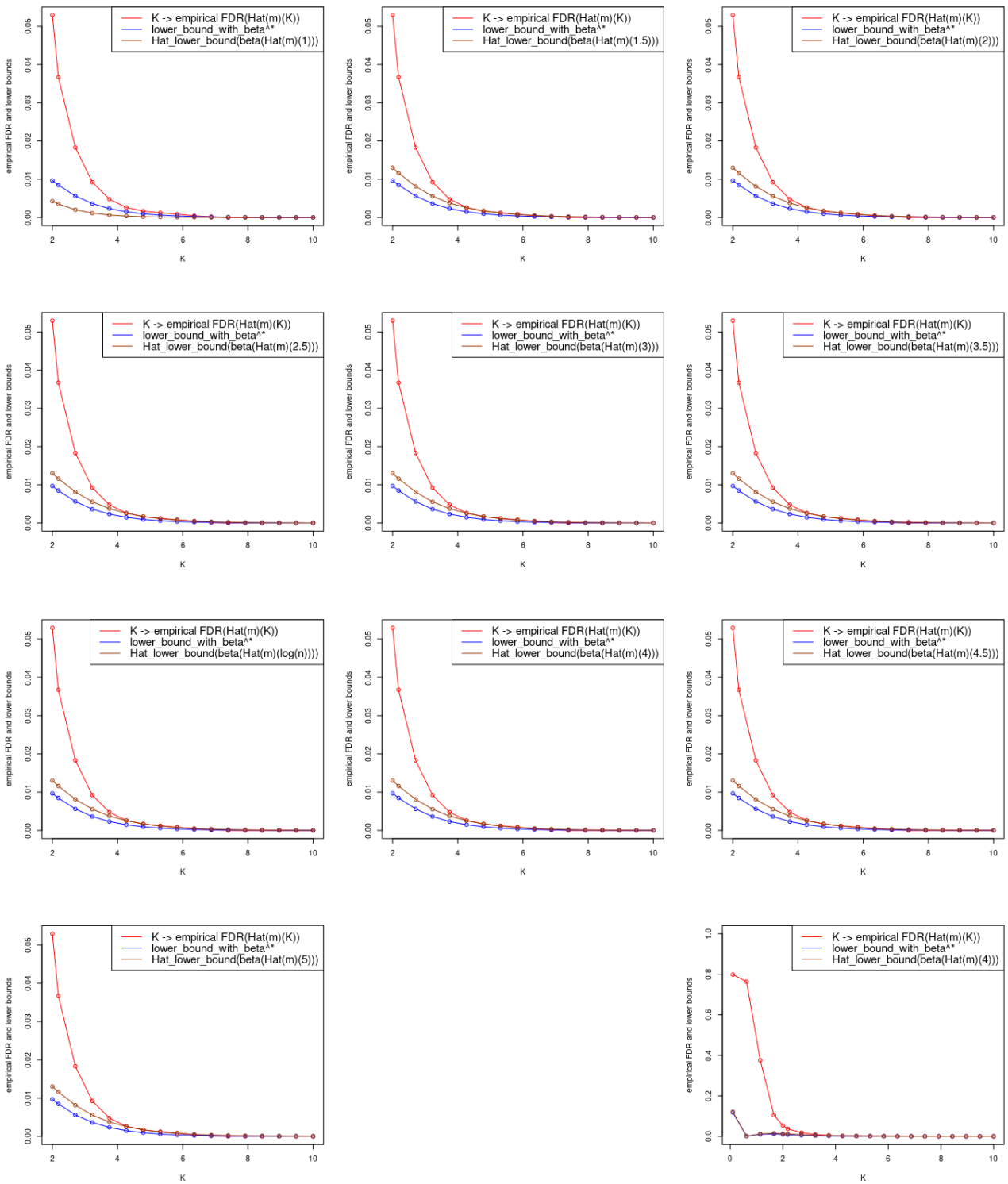


Figure 4.4: Comparison of the empirical estimation of the FDR, the function $b(K, \beta^*, \sigma^2)$ under the orthogonal design of Corollary 4.3 and the function $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$. The terms $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are calculating from only one dataset, independent of those used for the empirical estimations. For a better readability, we plot curves only for $K \geq 2$; but at the bottom right is the entire curve for $\tilde{K} = 4$.

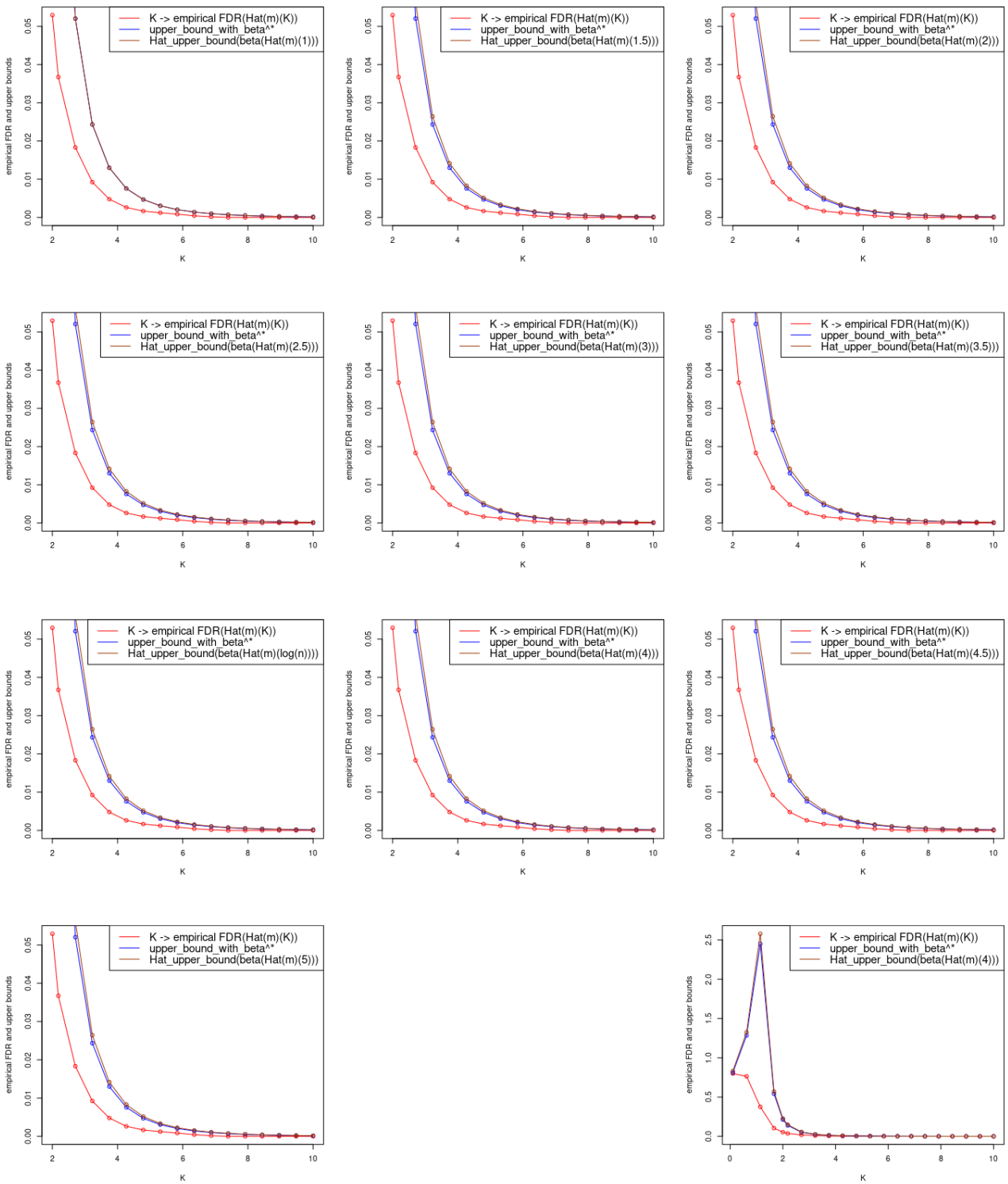


Figure 4.5: Comparison of the empirical estimation of the FDR, the function $B(K, \beta^*, \sigma^2)$ under the orthogonal design of Corollary 4.3 and the function $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$. The terms $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are calculating from only one dataset independent of those used for the empirical estimations. For a better readability, we plot curves only for $K \geq 2$; but at the bottom right is the entire curve for $\tilde{K} = 4$.

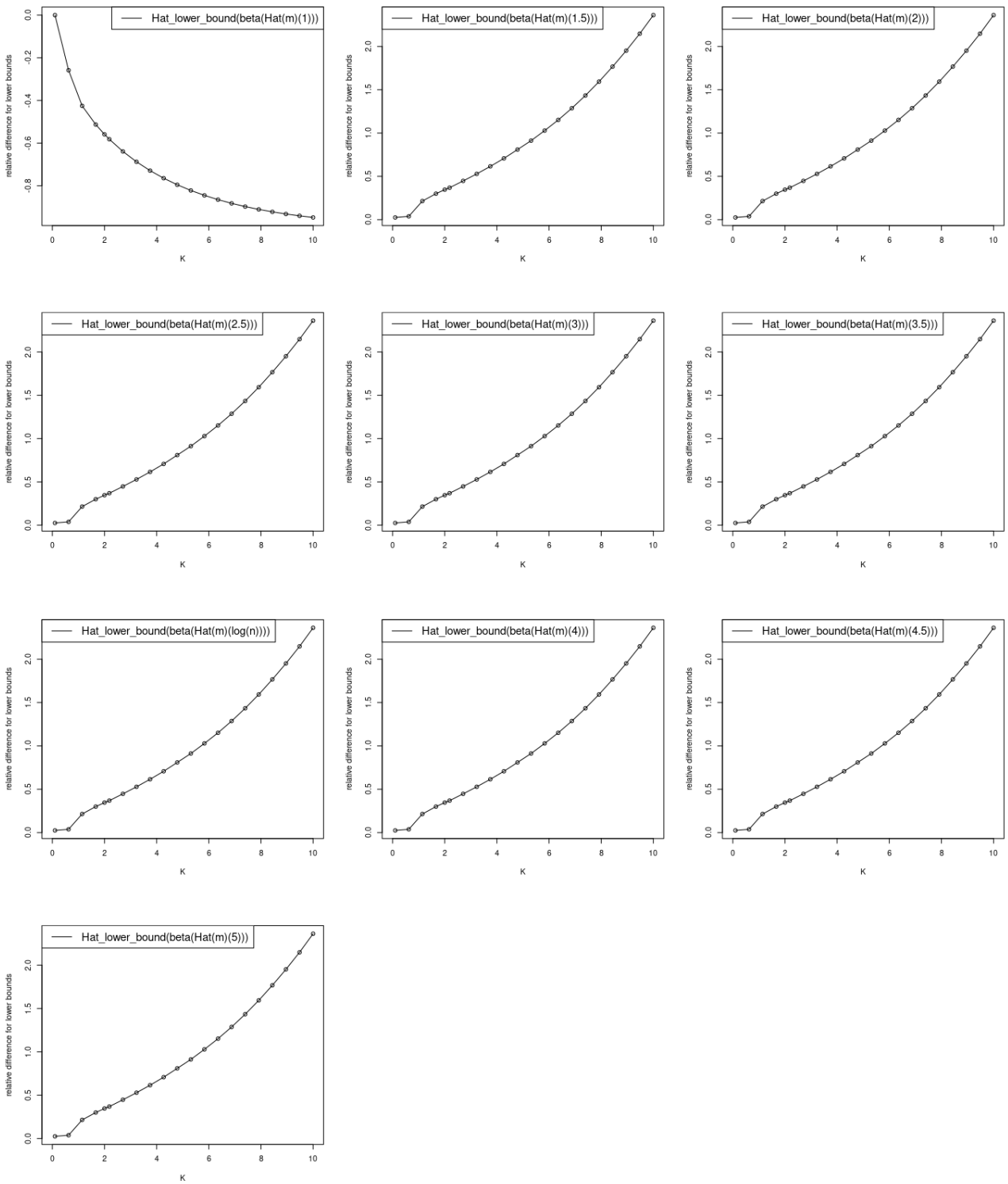


Figure 4.6: Curves of the relative change values between the function $b(K, \beta^*, \sigma^2)$ and the functions $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$, where estimators are calculating from only one dataset.

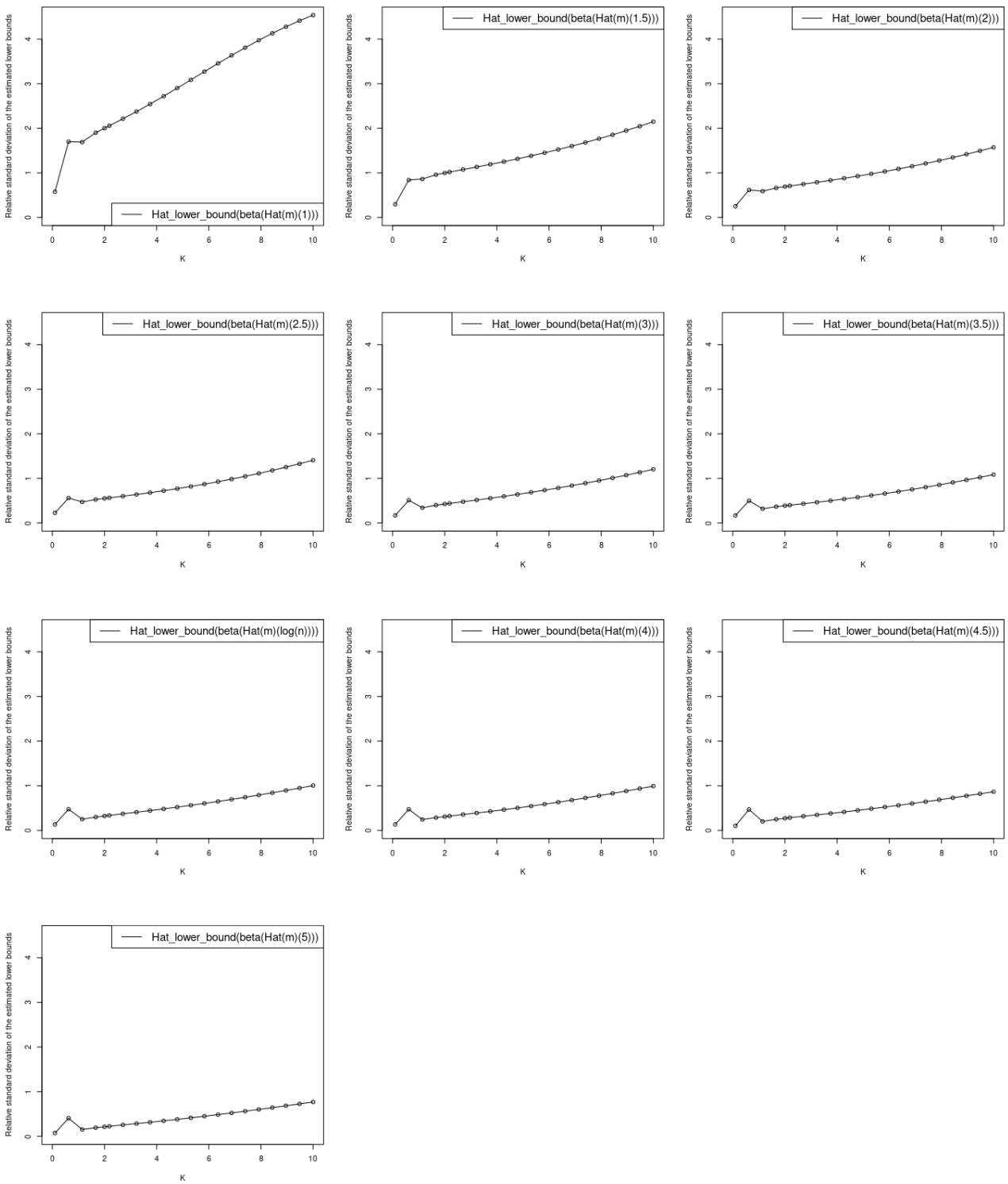


Figure 4.7: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $\tilde{b}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K.

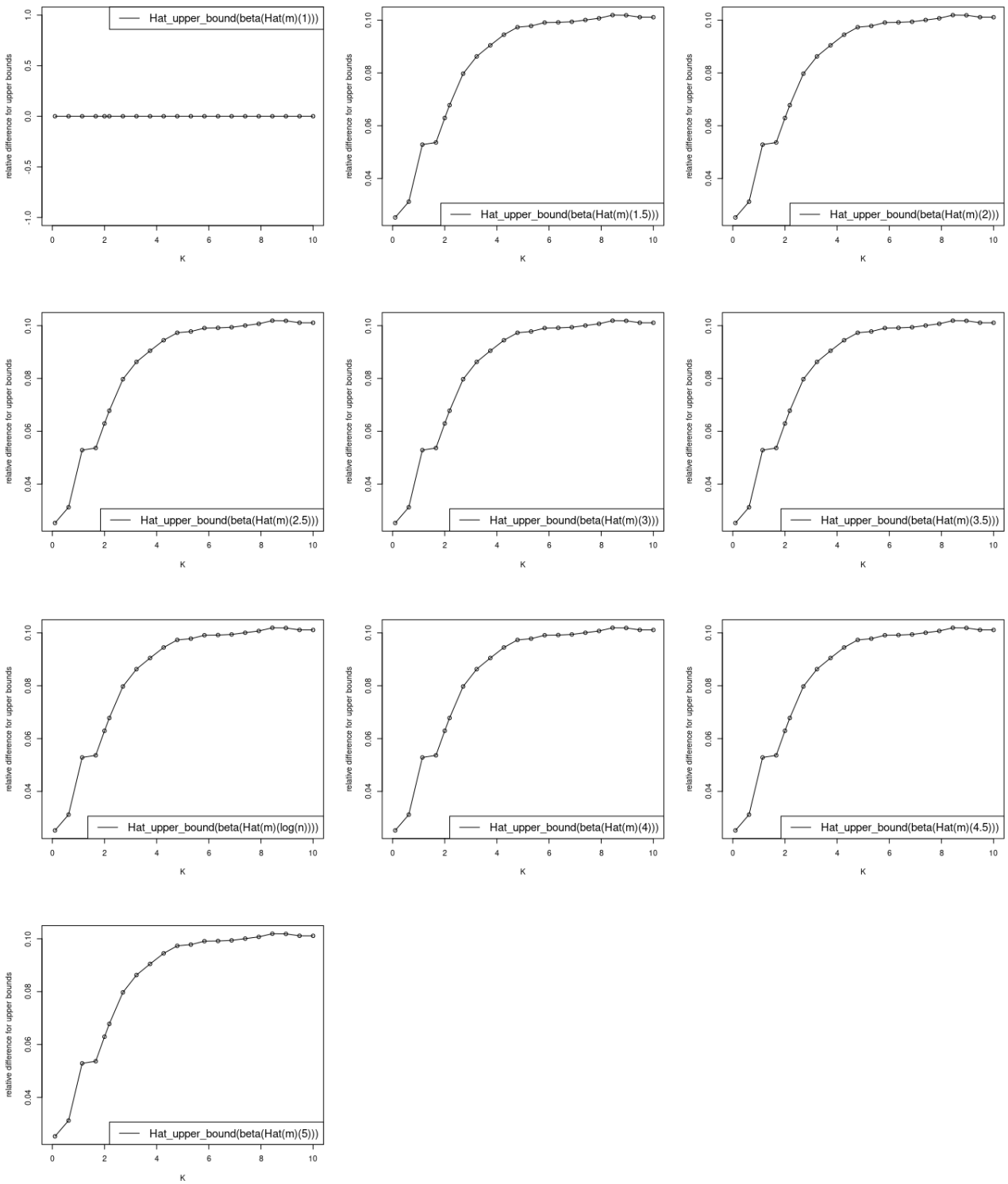


Figure 4.8: Curves of the relative change values between the function $B(K, \beta^*, \sigma^2)$ and the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$, where estimators are calculating from only one dataset.

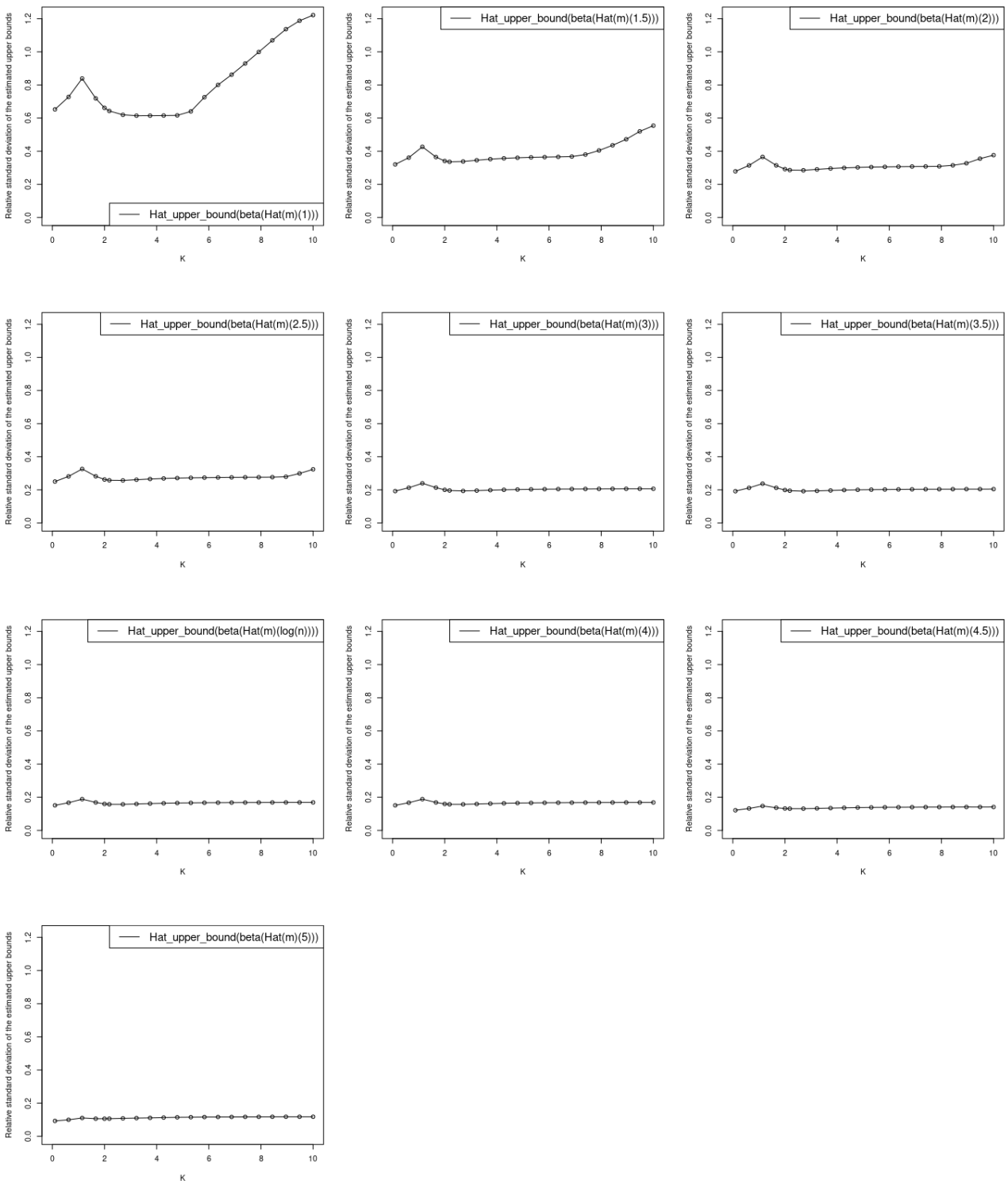


Figure 4.9: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K.

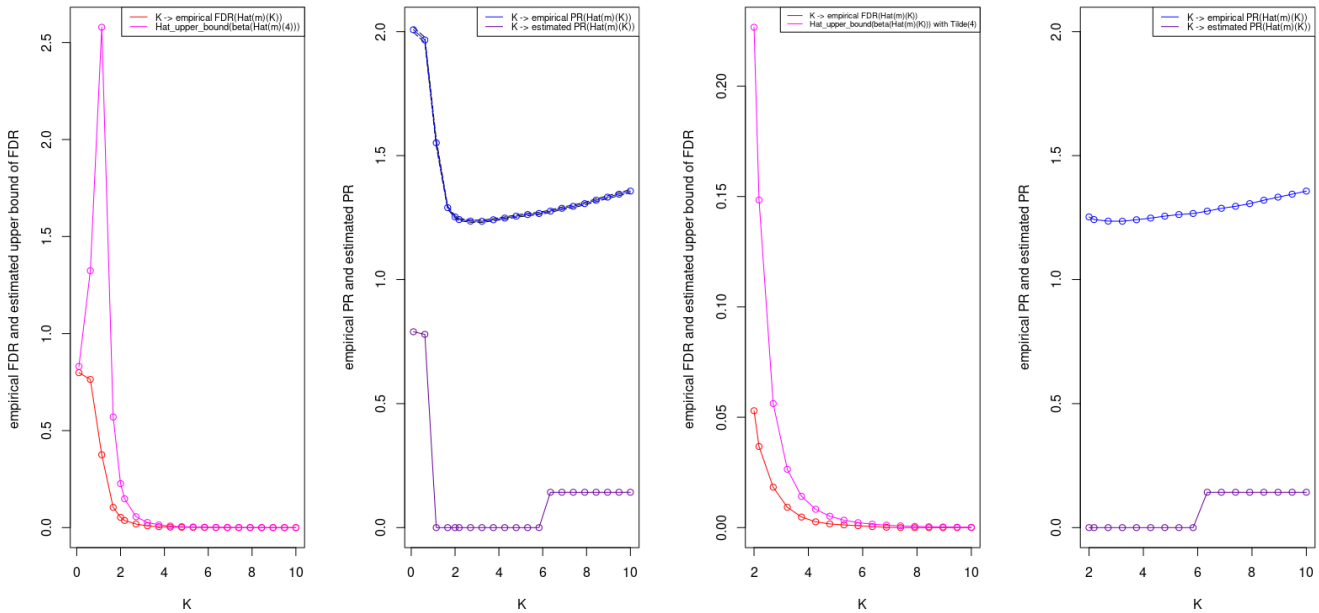


Figure 4.10: Left: Curves of the empirical estimation functions $\mathrm{FDR}\big(\hat{m}(K)\big)$ and $\mathrm{PR}\big(\hat{m}(K)\big)$ for all $K > 0$ by using 1000 datasets, and curves of the estimated risk (4.22) and the $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ function obtained in Corollary 4.3 by replacing β^* by $\hat{\beta}_{\hat{m}(4)}$. These two last plots are obtained from only one dataset. Right: curves are plotted only for $K \geq 2$.

4.5 Proofs.

4.5.1 FDR expression in model selection.

Proof of Formula 4.10.
If $D_m^* = q$, then $\mathrm{FP}(m) = 0$ for all $m \in \mathcal{M}$. So, $\mathrm{FDR}(m) = 0$ for all $m \in \mathcal{M}$; in particular for $m = \hat{m}$.
Let us now suppose that $D_m^* < q$.
The FDP expression within the model selection procedure is:

$$\begin{aligned}
\forall K > 0, \qquad \mathrm{FDP}(\hat{m}(K)) &= \frac{\mathrm{FP}(\hat{m}(K))}{D_{\hat{m}(K)} \vee 1} \\
&\underset{(*)}{=} \frac{D_{\hat{m}(K)} - D_{m^*}}{D_{\hat{m}(K)}} \mathbb{1}_{\{D_{\hat{m}(K)} > D_{m^*}\}} \\
&= \sum_{r=1}^{q} \frac{r - D_{m^*}}{r} \mathbb{1}_{\{r > D_{m^*}\}} \mathbb{1}_{\{D_{\hat{m}(K)} = r\}} \\
&\underset{(**)}{=} \sum_{r=D_{m^*}+1}^{q} \frac{r - D_{m^*}}{r} \mathbb{1}_{\{\hat{m}(K) = m_r\}} \\
&\underset{(***)}{=} \sum_{r=D_{m^*}+1}^{q} \frac{r - D_{m^*}}{r} \mathbb{1}_{\left\{\bigcap_{\substack{\ell=0 \\ \ell \neq r}}^{q} \{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\}\right\}}
\end{aligned} \tag{4.23}$$

(*) and *(**)* are due to the fact that models $(m)_{m \in \mathcal{M}}$ are nested and $m^* \in \mathcal{M}$. *(***)* is obtained since the crit_K function in injective on $\mathcal{M}$.
Finally, by taking the expectation in (4.23), we obtain the FDR expression (4.10). □

Proof of Proposition 4.1.
Before proving Proposition 4.1, let us cite and prove two lemmas.

Lemma 4.1. *For $r \in \{D_{m^*} + 1, \cdots, q\}$ and for all $\ell \in \{0, \cdots, r-1\}$:*

$$||Y - X\hat{\beta}_{m_r}||_2^2 - ||Y - X\hat{\beta}_{m_\ell}||_2^2 = -J_{1.r}$$

where $J_{1.r} = \sum_{k=\ell+1}^{r} \langle Y, u_k \rangle^2$

Lemma 4.2. *For $r \in \{D_{m^*} + 1, \cdots, q\}$ and for all $\ell \in \{r+1, \cdots, q\}$:*

$$||Y - X\hat{\beta}_{m_r}||_2^2 - ||Y - X\hat{\beta}_{m_\ell}||_2^2 = J_{2.r}$$

where $J_{2.r} = \sum_{k=r+1}^{\ell} \langle Y, u_k \rangle^2$

Proof of Lemma 4.1.
For $r \in \{D_{m^*}+1, \cdots, q\}$ and $\ell \in \{0, \cdots, r-1\}$:

$$\begin{aligned}
&||Y - X\hat{\beta}_{m_r}||_2^2 - ||Y - X\hat{\beta}_{m_\ell}||_2^2 = ||X\hat{\beta}_{m_r}||_2^2 - ||X\hat{\beta}_{m_\ell}||_2^2 + 2\langle Y, X\hat{\beta}_{m_\ell} - X\hat{\beta}_{m_r}\rangle \\
&= ||X\hat{\beta}_{m_r}||_2^2 - ||X\hat{\beta}_{m_\ell}||_2^2 + 2\langle Y - X\hat{\beta}_{m_r}, X\hat{\beta}_{m_\ell}\rangle \\
&\qquad\qquad - 2\langle Y - X\hat{\beta}_{m_r}, X\hat{\beta}_{m_r}\rangle + 2\langle X\hat{\beta}_{m_r}, X\hat{\beta}_{m_\ell}\rangle - 2||X\hat{\beta}_{m_r}||_2^2 \\
&= -||X\hat{\beta}_{m_r}||_2^2 - ||X\hat{\beta}_{m_\ell}||_2^2 + 2\langle X\hat{\beta}_{m_r}, X\hat{\beta}_{m_\ell}\rangle = -||X\hat{\beta}_{m_r} - X\hat{\beta}_{m_\ell}||_2^2
\end{aligned}$$

The last line is due to the fact that $Y - X\hat{\beta}_{m_r} \in (m_r)^\perp \subset (m_\ell)^\perp$ since $m_\ell \subset m_r$ and $X\hat{\beta}_{m_r}$ is the projection of Y onto m_r.
Then,

$$\begin{aligned}
||X\hat{\beta}_{m_r} - X\hat{\beta}_{m_\ell}||_2^2 &= ||\Pi_{m_r}(Y) - \Pi_{m_\ell}(Y)||_2^2 \\
&= ||\Pi_{\mathrm{Span}(X_1,\cdots,X_r)}(Y) - \Pi_{\mathrm{Span}(X_1,\cdots,X_\ell)}(Y)||_2^2 \\
&\underset{(*)}{=} ||\Pi_{\mathrm{Span}(u_1,\cdots,u_r)}(Y) - \Pi_{\mathrm{Span}(u_1,\cdots,u_\ell)}(Y)||_2^2 \\
&= ||\Pi_{\mathrm{Span}(u_{\ell+1},\cdots,u_r)}(Y)||_2^2 \\
&= ||\sum_{k=\ell+1}^{r} \langle Y, u_k\rangle u_k||_2^2 \\
&\underset{(**)}{=} \sum_{k=\ell+1}^{r} \langle Y, u_k\rangle^2
\end{aligned} \tag{4.24}$$

(*) come from the definition of $(u_1, \cdots, u_n)$ and *(**)* is obtained by Parseval equality.

□

Proof of Lemma 4.2.
For $r \in \{D_{m^*}+1, \cdots, q\}$ and $l \in \{r+1, \cdots, q\}$:

$$\begin{aligned}
||Y - X\hat{\beta}_{m_r}||_2^2 - ||Y - X\hat{\beta}_{m_\ell}||_2^2 &= ||X\hat{\beta}_{m_r}||_2^2 - ||X\hat{\beta}_{m_\ell}||_2^2 + 2\langle Y, X\hat{\beta}_{m_\ell} - X\hat{\beta}_{m_r}\rangle \\
&= ||X\hat{\beta}_{m_r}||_2^2 - ||X\hat{\beta}_{m_\ell}||_2^2 + 2\langle Y - X\hat{\beta}_{m_\ell}, X\hat{\beta}_{m_\ell}\rangle \\
&\qquad - 2\langle Y - X\hat{\beta}_{m_\ell}, X\hat{\beta}_{m_r}\rangle + 2||X\hat{\beta}_{m_\ell}||_2^2 - 2\langle X\hat{\beta}_{m_\ell}, X\hat{\beta}_{m_r}\rangle \\
&= ||X\hat{\beta}_{m_r}||_2^2 + ||X\hat{\beta}_{m_\ell}||_2^2 - 2\langle X\hat{\beta}_{m_\ell}, X\hat{\beta}_{m_r}\rangle \\
&= ||X\hat{\beta}_{m_\ell} - X\hat{\beta}_{m_r}||_2^2
\end{aligned}$$

The last line is due to the fact that $Y - X\hat{\beta}_{m_\ell} \in (m_\ell)^\perp \subset (m_r)^\perp$ since $m_r \subset m_\ell$, and $X\hat{\beta}_{m_\ell}$ is the projection of Y onto m_ℓ.

Then,

$$
\begin{aligned}
||X\hat{\beta}_{m_\ell} - X\hat{\beta}_{m_r}||_2^2 &= ||\Pi_{m_\ell}(Y) - \Pi_{m_r}(Y)||_2^2 \\
&= ||\Pi_{\mathrm{Span}(X_1,\cdots,X_\ell)}(Y) - \Pi_{\mathrm{Span}(X_1,\cdots,X_r)}(Y)||_2^2 \\
&\underset{(*)}{=} ||\Pi_{\mathrm{Span}(u_1,\cdots,u_\ell)}(Y) - \Pi_{\mathrm{Span}(u_1,\cdots,u_r)}(Y)||_2^2 \\
&= ||\Pi_{\mathrm{Span}(u_{r+1},\cdots,u_\ell)}(Y)||_2^2 \\
&= ||\sum_{k=r+1}^{\ell} \langle Y, u_k\rangle u_k||_2^2 \\
&\underset{(**)}{=} \sum_{k=r+1}^{\ell} \langle Y, u_k\rangle^2
\end{aligned}
\tag{4.25}
$$

(*) come from the definition of $(u_1, \cdots, u_n)$ and *(**)* is obtained by Parseval equality.

□

Proof of Proposition 4.1.

Starting from (4.10), we decompose the event $\left\{\bigcap_{\substack{\ell=0\\ \ell\neq r}}^{q}\{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\}\right\}$ by the intersection of these two events
$\left\{\bigcap_{\ell=0}^{r-1}\{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\}\right\}$ and $\left\{\bigcap_{\ell=r+1}^{q}\{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\}\right\}$.
By using the definition of the crit_K function, for $r \in \{D_{m^*}+1, \cdots, q\}$ and $\ell \in \{0, \cdots, r-1\}$, the event $\Big\{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\Big\}$ equals the event $\Big\{||Y - X\hat{\beta}_{m_r}||_2^2 - ||Y - X\hat{\beta}_{m_\ell}||_2^2 < K\sigma^2(\ell - r)\Big\}$ which is equal to $\Big\{-J_{1.r} < K\sigma^2(\ell - r)\Big\}$ according to (4.24).
Similarly, for $r \in \{D_{m^*}+1, \cdots, q\}$ and $\ell \in \{r+1, \cdots, q\}$, the event $\Big\{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\Big\}$ equals the event $\Big\{||Y - X\hat{\beta}_{m_r}||_2^2 - ||Y - X\hat{\beta}_{m_\ell}||_2^2 < K\sigma^2(\ell - r)\Big\}$ which is equal to $\Big\{J_{2.r} < K\sigma^2(\ell - r)\Big\}$ according to (4.25).
In this way, $\left\{\bigcap_{\substack{\ell=0\\ \ell\neq r}}^{q}\{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\}\right\}$ is decomposed by two events:
$\left\{\bigcap_{\ell=0}^{r-1}\left\{\sum_{k=\ell+1}^{r}\langle Y, u_k\rangle^2 > K\sigma^2(r-\ell)\right\}\right\} \cap \left\{\bigcap_{\ell=r+1}^{q}\left\{\sum_{k=r+1}^{\ell}\langle Y, u_k\rangle^2 < K\sigma^2(\ell - r)\right\}\right\}$
Let us define U the $n \times n$ matrix such that u_k is the $k-$th column of U. Since $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $(u_1, \cdots, u_n)$ is an orthonormal basis of $\mathbb{R}^n$, we get $U^T\varepsilon = \Big(\langle\varepsilon, u_1\rangle, \cdots, \langle\varepsilon, u_n\rangle\Big)^T \sim \mathcal{N}(0, \sigma^2 U I_n U^T) = \mathcal{N}(0, \sigma^2 I_n)$. Hence, $\langle Y, u_i\rangle_{i\in\{1,\cdots,n\}} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\langle X\beta^*, u_i\rangle, \sigma^2\big)$. Moreover, the first event depends only on $\langle Y, u_i\rangle_{i\in\{1,\cdots,r-1\}}$ whereas the second one depends only on $\langle Y, u_i\rangle_{i\in\{r+1,\cdots,q\}}$, so, the two events are independent.

Hence, from (4.10), we obtain for all $K > 0$:

$$\mathrm{FDR}(\hat{m}(K)) = \sum_{r=D_{m^*}+1}^{q} \frac{r - D_{m^*}}{r} \mathbb{P}\left(\bigcap_{\ell=0}^{r-1} \left\{ \sum_{k=\ell+1}^{r} \langle Y, u_k \rangle^2 > K\sigma^2(r-\ell) \right\} \right) \times \mathbb{P}\left(\bigcap_{\ell=r+1}^{q} \left\{ \sum_{k=r+1}^{\ell} \langle Y, u_k \rangle^2 < K\sigma^2(\ell - r) \right\} \right).$$

Moreover, since $\langle X\beta^*, u_k \rangle = 0, \forall k > D_{m^*}$ and since $r \geq D_m m^* + 1$, we have:

$$\sum_{k=\ell+1}^{r} \langle Y, u_k \rangle^2 = \sum_{k=\ell+1}^{r} \langle \varepsilon, u_k \rangle^2$$

So, for all $K > 0$ and for each $r \in \{D_{m^*} + 1, \cdots, q\}$:

$$\mathbb{P}\left(\bigcap_{\ell=r+1}^{q} \left\{ \sum_{k=r+1}^{\ell} \langle Y, u_k \rangle^2 < K\sigma^2(\ell - r) \right\} \right) = \mathbb{P}\Big(\bigcap_{\ell=r+1}^{q} \Big\{ \sum_{k=r+1}^{\ell} \tilde{Z}_k^{\,2} < K\sigma^2(\ell - r) \Big\} \Big),$$

$$\text{where } \tilde{Z}_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\mathbb{P}\left(\bigcap_{\ell=r+1}^{q} \left\{ \sum_{k=r+1}^{\ell} \langle Y, u_k \rangle^2 < K\sigma^2(\ell - r) \right\} \right) = \mathbb{P}\Big(\bigcap_{\ell=r+1}^{q} \Big\{ \sum_{k=r+1}^{\ell} Z_k^2 < K(\ell - r) \Big\} \Big),$$

$$\text{where } Z_k \overset{i.i.d.}{\sim} \mathcal{N}(0, 1).$$

Hence, for all $K > 0$ and for each $r \in \{D_{m^*}+1, \cdots, q\}$, $\mathbb{P}\left(\bigcap_{\ell=r+1}^{q} \left\{ \sum_{k=r+1}^{\ell} \langle Y, u_k \rangle^2 < K\sigma^2(\ell - r) \right\} \right)$ does not depend on the data at all and we deduce the Formula (4.11) with:

$$P_r(K) = \mathbb{P}\Big(\bigcap_{\ell=r+1}^{q} \Big\{ \sum_{k=r+1}^{\ell} Z_k^2 < K(\ell - r) \Big\} \Big),$$

$$Q_r(K, \beta^*, \sigma^2) = \mathbb{P}\Big(\bigcap_{\ell=0}^{r-1} \Big\{ \sum_{k=\ell+1}^{r} \langle Y, u_k \rangle^2 > K\sigma^2(r - \ell) \Big\} \Big),$$

where $Z_k \overset{i.i.d.}{\sim} \mathcal{N}(0, 1),\ \forall k \in \{r + 1, \cdots, q\}$. □

4.5.2 General bounds.

Proof of Theorem 4.1.
We start from (4.11).

- ***Bounds of the Q_r terms.***
For all $K > 0$ and for each $r \in \{D_{m^*} + 1, \cdots, q\}$:

$$Q_r(K, \beta^*, \sigma^2) = \mathbb{P}\left(\bigcap_{\ell=0}^{r-1} \left\{ \sum_{k=\ell+1}^{r} \langle Y, u_k \rangle^2 > K\sigma^2(r-\ell) \right\}\right)$$

Moreover, since $\langle X\beta^*, u_k \rangle = 0, \forall k > D_{m^*}$, we have:

$$\sum_{k=\ell+1}^{r} \langle Y, u_k \rangle^2 = \sum_{k=\ell+1}^{r} \left(\langle \varepsilon, u_k \rangle^2 \mathbb{1}_{k > D_{m^*}} + \langle Y, u_k \rangle^2 \mathbb{1}_{k \leq D_{m^*}} \right) \tag{4.26}$$

- - **Lower bound of $Q_r(K, \beta^*, \sigma^2)$ for all $K > 0$ and for each $r \in \{D_{m^*} + 1, \cdots, q\}$:**

Lemma 4.3. *Let us consider an integer $s > 1$, $K > 0$ and $c_1, \cdots, c_s$ s positive random independent quantities. We define by E_ℓ the event $\{c_\ell > \ell K\sigma^2\}$ for $\ell \in \{1, \cdots, s\}$ and by F_ℓ the event $\{K\sigma^2 < c_\ell \leq \ell K\sigma^2\}$ for ℓ in $\{2, \cdots, s\}$.*
Then:

$$\begin{aligned}
&\left\{c_s > K\sigma^2\right\} \cap \left\{c_s + c_{s-1} > 2K\sigma^2\right\} \cap \cdots \cap \left\{c_s + c_{s-1} + \cdots + c_1 > sK\sigma^2\right\} \\
&\supseteq E_s \sqcup \left(F_s \sqcap \left(E_{s-1} \sqcup \left(F_{s-1} \sqcap \left(E_{s-2} \sqcup \cdots \sqcup (F_3 \sqcap (E_2 \sqcup (F_2 \sqcap E_1)))\right)\right)\right)\right)
\end{aligned}$$

where $\cap$ and $\sqcap$ design respectively any intersection and a disjoint intersection of events, as well as $\cup$ and $\sqcup$ designing respectively any union and a disjoint union of events.

Proof. We prove Lemma 4.3 by a recurrence on $s \geq 1$.

The inclusion is true for $s = 1$. Let $s \geq 1$ and suppose that the inclusion is true for s, then:

$$\begin{aligned}
&\left\{c_{s+1} > K\sigma^2\right\} \cap \left\{c_{s+1} + c_s > 2K\sigma^2\right\} \cap \cdots \cap \left\{c_{s+1} + c_s + \cdots + c_1 > (s+1)K\sigma^2\right\} \\
&= \left(E_{s+1} \sqcup F_{s+1}\right) \cap \left(\left\{c_{s+1} + c_s > 2K\sigma^2\right\} \cap \cdots \cap \left\{c_{s+1} + c_s + \cdots + c_1 > (s+1)K\sigma^2\right\}\right) \\
&= \left(E_{s+1} \cap \left(\left\{c_{s+1} + c_s > 2K\sigma^2\right\} \cap \cdots \cap \left\{c_{s+1} + c_s + \cdots + c_1 > (s+1)K\sigma^2\right\}\right)\right) \\
&\quad \sqcup \left(F_{s+1} \cap \left(\left\{c_{s+1} + c_s > 2K\sigma^2\right\} \cap \cdots \cap \left\{c_{s+1} + c_s + \cdots + c_1 > (s+1)K\sigma^2\right\}\right)\right) \\
&\underset{(*)}{=} E_{s+1} \sqcup \left(F_{s+1} \cap \left(\left\{c_{s+1} + c_s > 2K\sigma^2\right\} \cap \cdots \cap \left\{c_{s+1} + c_s + \cdots + c_1 > (s+1)K\sigma^2\right\}\right)\right) \\
&\supseteq E_{s+1} \sqcup \left(F_{s+1} \cap \left(\left\{c_s > K\sigma^2\right\} \cap \left\{c_s + c_{s-1} > 2K\sigma^2\right\} \cap \cdots \cap \left\{c_s + c_{s-1} + \cdots + c_1 > sK\sigma^2\right\}\right)\right) \\
&\underset{(**)}{\supseteq} E_{s+1} \sqcup \left(F_{s+1} \cap \left(E_s \sqcup \left(F_s \sqcap \left(E_{s-1} \sqcup \cdots \sqcup \left(F_3 \sqcap \left(E_3 \sqcup \left(F_2 \sqcap E_1\right)\right)\right)\right)\right)\right)\right) \\
&\underset{(***)}{\supseteq} E_{s+1} \sqcup \left(F_{s+1} \sqcap \left(E_s \sqcup \left(F_s \sqcap \left(E_{s-1} \sqcup \cdots \sqcup \left(F_3 \sqcap \left(E_3 \sqcup \left(F_2 \sqcap E_1\right)\right)\right)\right)\right)\right)\right)
\end{aligned}$$

(*) is true since $E_{s+1} \subset \left(\left\{c_{s+1} + c_s > 2K\sigma^2\right\} \cap \cdots \cap \left\{c_{s+1} + c_s + \cdots + c_1 > (s+1)K\sigma^2\right\}\right)$, *(**)* is obtained by applying the recurrence assumption at the step s and the independence between F_{s+1} and $\left(E_s \sqcup \left(F_s \sqcap \left(E_{s-1} \sqcup \cdots \sqcup \left(F_3 \sqcap \left(E_3 \sqcup \left(F_2 \sqcap E_1\right)\right)\right)\right)\right)\right)$ gives *(***)*.
Thus, the property is true for $s + 1$, which proves lemma. $\square$

By applying Lemma 4.3 with $s = r$, $c_\ell = \langle \varepsilon, u_\ell \rangle^2$ $\forall \ell \in \{D_{m^*} + 1, \cdots, r\}$ and $c_\ell = \langle Y, u_\ell \rangle^2$ $\forall \ell \in$

$\{1, \cdots D_{m^*}\}$, we obtain, by using the notation of Lemma 4.3 and formula (4.26),

$$\begin{aligned}
Q_r(K, \beta^*, \sigma^2) &= \mathbb{P}(\Big\{\langle \varepsilon, u_r\rangle^2 > K\sigma^2\Big\} \cap \cdots \cap \Big\{\langle \varepsilon, u_r\rangle^2 + \cdots + \langle \varepsilon, u_{D_{m^*}+1}\rangle^2 > K\sigma^2(r - D_{m^*})\Big\} \\
&\cap \Big\{\langle \varepsilon, u_r\rangle^2 + \cdots + \langle \varepsilon, u_{D_{m^*}+1}\rangle^2 + \langle Y, u_{D_{m^*}}\rangle^2 > K\sigma^2(r - D_{m^*} + 1)\Big\} \cap \cdots \\
&\cap \Big\{\langle \varepsilon, u_r\rangle^2 + \cdots + \langle \varepsilon, u_{D_{m^*}+1}\rangle^2 + \langle Y, u_{D_{m^*}}\rangle^2 + \cdots + \langle Y, u_1\rangle^2 > K\sigma^2 r\Big\} \\
&= \mathbb{P}\Bigg(\Big\{c_r > K\sigma^2\Big\} \cap \Big\{c_r + c_{r-1} > 2K\sigma^2\Big\} \cap \cdots \cap \Big\{c_r + c_{r-1} + \cdots + c_1 > rK\sigma^2\Big\}\Bigg) \\
&\geq \mathbb{P}(E_r) + \mathbb{P}(F_r)\Bigg(\mathbb{P}(E_{r-1}) + \mathbb{P}(F_{r-1})\Big(\mathbb{P}(E_{r-2}) + \cdots + \mathbb{P}(F_3)\Big(\mathbb{P}(E_2) + \mathbb{P}(F_2)\mathbb{P}(E_1)\Big)\Big)\Bigg)
\end{aligned} \tag{4.27}$$

Since $\langle \varepsilon, u_\ell\rangle$ follows a centered Gaussian distribution with a variance σ^2 $\forall \ell \in \{1, \cdots, r\}$ and $\langle Y, u_\ell\rangle$ follows a Gaussian distribution with mean $\langle X\beta^*, u_\ell\rangle$ $\forall \ell \in \{1, \cdots, D_{m^*}\}$ and variance σ^2, we get:

$$\begin{aligned}
\text{For } \ell \in \{D_{m^*} + 1, \cdots, r\}: \quad \mathbb{P}(E_\ell) &= \mathbb{P}\Big(\Big\{c_\ell > \ell K\sigma^2\Big\}\Big) \\
&= 2\Big(1 - \Phi\big(\sqrt{\ell K}\big)\Big) = G_\ell \\
\mathbb{P}(F_\ell) &= \mathbb{P}\Big(\Big\{K\sigma^2 < c_\ell \leq \ell K\sigma^2\Big\}\Big) \\
&= 2\Big(\Phi\big(\sqrt{\ell K}\big) - \Phi\big(\sqrt{K}\big)\Big) = H_\ell
\end{aligned}$$

$$\begin{aligned}
\text{For } \ell \in \{1, \cdots, D_{m^*}\}: \quad \mathbb{P}(E_\ell) &= \mathbb{P}\Big(\Big\{c_\ell > \ell K\sigma^2\Big\}\Big) \\
&= 2 - \Bigg(\Phi\Big(\sqrt{\ell K} - \frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big) + \Phi\Big(\sqrt{\ell K} + \frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big)\Bigg) \\
&= G_\ell
\end{aligned} \tag{4.28}$$

$$\begin{aligned}
\text{For } \ell \in \{2, \cdots, D_{m^*}\}: \quad \mathbb{P}(F_\ell) &= \mathbb{P}\Big(\Big\{K\sigma^2 < c_\ell \leq \ell K\sigma^2\Big\}\Big) \\
&= \Phi\Big(\sqrt{\ell K} - \frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big) + \Phi\Big(\sqrt{\ell K} + \frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big) \qquad (4.29) \\
&- \Bigg(\Phi\Big(\sqrt{K} - \frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big) + \Phi\Big(\sqrt{K} + \frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big)\Bigg) \\
&= H_\ell \qquad (4.30)
\end{aligned}$$

Hence, a lower bound of $Q_r(K, \beta^*, \sigma^2)$ is obtained for all $K > 0$:

$$\underline{f}_r(K, \beta^*, \sigma^2) \leq Q_r(K, \beta^*, \sigma^2) \tag{4.31}$$

with:

$$\begin{aligned} \underline{f}_r(K, \beta^*, \sigma^2) &= G_r + H_r \, \underline{f}_{r-1}(K, \beta^*) \\ \text{with} \quad \underline{f}_1(K, \beta^*) &= G_1 \end{aligned} \tag{4.32}$$

- - **Upper bound of $Q_r(K, \beta^*, \sigma^2)$ for all $K > 0$ and for each $r \in \{D_{m^*} + 1, \cdots, q\}$:**
Denote by $c_\ell = \langle Y, u_\ell \rangle^2 \; \forall \ell \in \{1, \cdots, D_{m^*}\}$ and by $c_\ell = \langle \varepsilon, u_\ell \rangle^2 \; \forall \ell \in \{D_{m^*} + 1, r\}$, we have:

$$\begin{aligned} Q_r(K, \beta^*, \sigma^2) &= \mathbb{P}\Big(\Big\{ c_r > K\sigma^2 \Big\} \cap \Big\{ c_r + c_{r-1} > 2K\sigma^2 \Big\} \cap \cdots \cap \Big\{ c_r + c_{r-1} + \cdots + c_1 > rK\sigma^2 \Big\} \Big) \\ &\leq \min \Bigg(\mathbb{P}\Big(\Big\{ c_r > K\sigma^2 \Big\} \Big), \mathbb{P}\Big(\Big\{ c_r + c_{r-1} > 2K\sigma^2 \Big\} \Big), \cdots, \mathbb{P}\Big(\Big\{ c_r + c_{r-1} + \cdots + c_1 > rK\sigma^2 \Big\} \Big) \Bigg) \end{aligned} \tag{4.33}$$

For all $j \in \{D_{m^*} + 1, r\}$,

$$\begin{aligned} \mathbb{P}\Big(\Big\{ c_r + \cdots + c_j > (r - j + 1)K\sigma^2 \Big\} \Big) &\underset{(*)}{=} \mathbb{P}\Big(\Big\{ X > (r - j + 1)K \Big\} \Big) \quad \text{for } X \sim \chi^2(r - j + 1) \\ &= 1 - F_{\chi^2(r-j+1)}\Big((r - j + 1)K \Big). \end{aligned} \tag{4.34}$$

(*) is true since $\langle \varepsilon, u_i \rangle_{i \in \{D_{m^*}+1, \cdots, r\}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, \sigma^2\big)$.

For all $j \in \{1, \cdots, D_{m^*}\}$,

$$
\begin{aligned}
&\mathbb{P}\Big(\Big\{c_r + \cdots + c_j > (r-j+1)K\sigma^2\Big\}\Big) \\
&= \mathbb{P}\Big(\Big\{c_r + \cdots + c_{D_{m^*}+1} + c_{D_{m^*}} + \cdots + c_j > (r-j+1)K\sigma^2\Big\}\Big) \\
&= \mathbb{P}\Big(\Big\{c_r + \cdots + c_{D_{m^*}+1} + \big(\langle X\beta^*, u_{D_{m^*}}\rangle + \langle \varepsilon, u_{D_{m^*}}\rangle\big)^2 + \cdots + \big(\langle X\beta^*, u_j\rangle + \langle \varepsilon, u_j\rangle\big)^2 \\
&\qquad\qquad > (r-j+1)K\sigma^2\Big\}\Big) \\
&\underset{(**)}{\leq} \mathbb{P}\Big(\Big\{c_r + \cdots + c_{D_{m^*}+1} + 2\langle X\beta^*, u_{D_{m^*}}\rangle^2 + 2\langle \varepsilon, u_{D_{m^*}}\rangle^2 + \cdots \\
&\qquad\qquad + 2\langle X\beta^*, u_j\rangle^2 + 2\langle \varepsilon, u_j\rangle^2 > (r-j+1)K\sigma^2\Big\}\Big) \\
&\leq \mathbb{P}\Big(\Big\{2c_r + \cdots + 2c_{D_{m^*}+1} + 2\langle \varepsilon, u_{D_{m^*}}\rangle^2 + \cdots + 2\langle \varepsilon, u_j\rangle^2 > (r-j+1)K\sigma^2 \\
&\qquad\qquad - 2\langle X\beta^*, u_{D_{m^*}}\rangle^2 - \cdots - 2\langle X\beta^*, u_j\rangle^2\Big\}\Big) \\
&\underset{(***)}{=} \mathbb{P}\Big(\Big\{2\sigma^2 Z_r^2 + \cdots + 2\sigma^2 Z_{D_{m^*}+1}^2 + 2\sigma^2 Z_{D_{m^*}}^2 + \cdots + 2\sigma^2 Z_j^2 \\
&> (r-j+1)K\sigma^2 - 2\langle X\beta^*, u_{D_{m^*}}\rangle^2 - \cdots - 2\langle X\beta^*, u_j\rangle^2\Big\}\Big), \quad \text{where } (Z_\ell)_{\ell\in\{j,\cdots,r\}} \overset{i.i.d}{\sim} \mathcal{N}(0,1). \\
&= \mathbb{P}\Big(\Big\{Z_r^2 + \cdots + Z_{D_{m^*}+1}^2 + Z_{D_{m^*}}^2 + \cdots + Z_j^2 > \frac{(r-j+1)K}{2} - \frac{\langle X\beta^*, u_{D_{m^*}}\rangle^2}{\sigma^2} - \cdots - \frac{\langle X\beta^*, u_j\rangle^2}{\sigma^2}\Big\}\Big) \\
&= \mathbb{P}\Big(\Big\{X > \frac{(r-j+1)K}{2} - \frac{\langle X\beta^*, u_{D_{m^*}}\rangle^2}{\sigma^2} - \cdots - \frac{\langle X\beta^*, u_j\rangle^2}{\sigma^2}\Big\}\Big), \quad \text{for } X \sim \chi^2(r-j+1) \\
&= 1 - F_{\chi^2(r-j+1)}\Big(\frac{(r-j+1)K}{2} - \frac{\langle X\beta^*, u_{D_{m^*}}\rangle^2}{\sigma^2} - \cdots - \frac{\langle X\beta^*, u_j\rangle^2}{\sigma^2}\Big). \qquad (4.35)
\end{aligned}
$$

(**) provides from $(a+b)^2 \leq 2(a^2+b^2),\ \forall (a,b) \in \mathbb{R}$ and *(***)* is true since $\langle \varepsilon, u_i\rangle_{i\in\{1,\cdots,r\}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0,\sigma^2\big)$.

So, from (4.33), (4.34) and (4.35), we deduce that for all $K > 0$ and for each $r \in \{D_{m^*} +$

$1, \cdots, q\}$:

$$Q_r(K, \beta^*, \sigma^2) \leq \min \Bigg(1 - F_{\chi^2(1)}(K), \cdots, 1 - F_{\chi^2(r-D_{m^*})}\big((r - D_{m^*})K\big),$$
$$1 - F_{\chi^2(r-D_{m^*}+1)}\left(\frac{(r - D_{m^*} + 1)K}{2} - \frac{\langle X\beta^*, u_{D_{m^*}}\rangle^2}{\sigma^2}\right),$$
$$1 - F_{\chi^2(r-D_{m^*}+2)}\left(\frac{(r - D_{m^*} + 2)K}{2} - \frac{\langle X\beta^*, u_{D_{m^*}}\rangle^2}{\sigma^2} - \frac{\langle X\beta^*, u_{D_{m^*-1}}\rangle^2}{\sigma^2}\right),$$
$$\cdots,$$
$$1 - F_{\chi^2(r)}\left(\frac{rK}{2} - \frac{\langle X\beta^*, u_{D_{m^*}}\rangle^2}{\sigma^2} - \frac{\langle X\beta^*, u_{D_{m^*-1}}\rangle^2}{\sigma^2} - \cdots - \frac{\langle X\beta^*, u_1\rangle^2}{\sigma^2}\right)\Bigg)$$

Hence, an upper bound of $Q_r(K, \beta^*, \sigma^2)$ is obtained for all $K > 0$:

$$Q_r(K, \beta^*, \sigma^2) \leq \overline{f}_r(K, \beta^*, \sigma^2)) \tag{4.36}$$

with:

$$\overline{f}_r(K, \beta^*, \sigma^2)) = 1 - \max\left(\max_{\ell \in \{1, \cdots, r-D_{m^*}\}} \Big(F_{\chi^2(\ell)}(\ell K)\Big), \max_{\ell \in \{r-D_{m^*}+1, r\}} \left(F_{\chi^2(\ell)}\Big(\frac{\ell K}{2} - \sum_{k=r-\ell+1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}\Big)\right)\right) \tag{4.37}$$

- ***Bounds of the FDR.***

By combining (4.11), (4.31), (4.32), (4.36), (4.37) and (4.12), we obtain:

$$\sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} P_r(K) \underline{f}_r(K, \beta^*, \sigma^2)\right) \leq \mathrm{FDR}(\hat{m}(K)) \leq \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} P_r(K) \overline{f}_r(K, \beta^*, \sigma^2))\right)$$

which allows us to obtain Theorem 4.1 with $\forall K > 0$,

$$b(K, \beta^*, \sigma^2) = \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} P_r(K) \underline{f}_r(K, \beta^*, \sigma^2)\right)$$

and

$$B(K, \beta^*, \sigma^2) = \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} P_r(K) \overline{f}_r(K, \beta^*, \sigma^2))\right)$$

□

4.5.3 In a no noise framework, FDR is strictly positive.

Proof of Corollary 4.1.

From Theorem 4.1, we have $\forall K > 0$,

$$\mathrm{FDR}(\hat{m}(K)) \geq \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} P_r(K)\ \underline{f}_r(K, \beta^*, \sigma^2)\right) \tag{4.38}$$

For all $K > 0$, $P_r(K)$ is non zero since the right parts of the inequality are always strictly positive making each considered involved event possible. For the rest of the proof, we use the following Lemma from Frank R.Kschischang:

Lemma 4.4 (Frank R. Kschischang [Kschischang, 2017])**.** *The complementary error function, erfc(x), is defined, for $x \geq 0$, as:*

$$erfc(x) = 2\Big(1 - F_{\mathcal{N}(0,\frac{1}{2})}(x)\Big)$$

where $F_{\mathcal{N}(0,\frac{1}{2})}$ designs the cumulative function of the centered Gaussian with the variance equals $\frac{1}{2}$.
Then,

$$\forall x \geq 0, \qquad \frac{2e^{-x^2}}{\sqrt{\pi}\big(x + \sqrt{x^2+2}\big)} \leq erfc(x) \leq \frac{e^{-x^2}}{\sqrt{\pi}x} \tag{4.39}$$

We remark that for all $x \geq 0$, $1 - \Phi(x) = \frac{1}{2}\text{erfc}\big(\frac{x}{\sqrt{2}}\big)$. Then, for each $r \in \{D_{m^*}+1, \cdots, q\}$,

$$\begin{aligned}
\underline{f}_r(K, \beta^*, \sigma^2) &= G_r + H_r\Big(G_{r-1} + H_{r-1}\big(G_{r-2} + \cdots + H_2 G_1\big)\Big) \\
&\geq G_r \\
&= 2\Big(1 - \Phi\big(\sqrt{rK}\big)\Big) \\
&= \text{ercf}\Big(\frac{\sqrt{rK}}{\sqrt{2}}\Big) \\
&\underset{(**)}{\geq} \frac{2}{\sqrt{\pi}\Big(\frac{\sqrt{rK}}{\sqrt{2}} + \sqrt{\frac{rK}{2}+2}\Big)} e^{-\frac{rK}{2}} \\
&= \frac{2\sqrt{2}}{\sqrt{\pi}\Big(\sqrt{rK} + \sqrt{rK+4}\Big)} e^{-\frac{rK}{2}}
\end{aligned} \tag{4.40}$$

(**) is provided by (4.39). So, from (4.38) and (4.40), we obtain:

$$\forall K > 0, \quad \text{FDR}(\hat{m}(K)) \geq \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} P_r(K) \frac{2\sqrt{2}}{\sqrt{\pi}\Big(\sqrt{rK} + \sqrt{rK+4}\Big)} e^{-\frac{rK}{2}} \right)$$

This lower bound is strictly positive and so is the FDR function. □

4.5.4 Asymptotic analysis.

Proof of Corollary 4.2.
For all $r \in \{D_{m^*}+1, \cdots, q\}$ and by using the definitions from Theorem 4.1,

$$\text{for } \ell \in \{D_{m^*}+1, \cdots, r\}: \; G_\ell = 2\Big(1-\Phi\big(\sqrt{\ell K}\big)\Big) \underset{K\longrightarrow+\infty}{\longrightarrow} 0$$

$$H_\ell = 2\Big(\Phi\big(\sqrt{\ell K}\big)-\Phi\big(\sqrt{K}\big)\Big) \underset{K\longrightarrow+\infty}{\longrightarrow} 0$$

$$\text{for } \ell \in \{1, \cdots, D_{m^*}\}: \; G_\ell = 2-\left(\Phi\Big(\sqrt{\ell K}-\frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\rangle\Big)+\Phi\Big(\sqrt{\ell K}+\frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big)\right) \underset{K\longrightarrow+\infty}{\longrightarrow} 0$$

$$\text{for } \ell \in \{2, \cdots, D_{m^*}\}: \; H_\ell = \Phi\Big(\sqrt{\ell K}-\frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big)+\Phi\Big(\sqrt{\ell K}+\frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big)-$$
$$\left(\Phi\Big(\sqrt{K}-\frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big)+\Phi\Big(\sqrt{K}+\frac{\langle X\beta^*, u_\ell\rangle}{\sigma}\Big)\right) \underset{K\longrightarrow+\infty}{\longrightarrow} 0$$

which leads to $\underline{f}_r(K, \beta^*, \sigma^2) \underset{K\longrightarrow+\infty}{\longrightarrow} 0$.

Moreover, $\overline{f}_r(K, \beta^*, \sigma^2)) = 1-\max\Bigg(\max\limits_{\ell\in\{1,\cdots,r-D_{m^*}\}}\Big(F_{\chi^2(\ell)}(\ell K)\Big),$
$\max\limits_{\ell\in\{r-D_{m^*}+1,\cdots,r\}}\Bigg(F_{\chi^2(\ell)}\Big(\frac{\ell K}{2}-\sum\limits_{k=r-\ell+1}^{D_{m^*}}\frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}\Big)\Bigg)\Bigg)$ tends to 0 when K tends to $+\infty$. We deduce that $Q_r(K, \beta^*, \sigma^2) \underset{K\longrightarrow+\infty}{\longrightarrow} 0$.

In the same way, for all $r \in \{D_{m^*}+1, \cdots, q\}$, $P_r(K) = \mathbb{P}\Big(\bigcap\limits_{\ell=r+1}^{q}\Big\{\sum\limits_{k=r+1}^{\ell} Z_k^2 < K(\ell-r)\Big\}\Big)$, where $Z_k \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \forall k \in \{r+1, \cdots, q\}$ tends to 1 when $K \longrightarrow +\infty$. Finally, since $P_r(K)Q_r(K, \beta^*, \sigma^2) \underset{K\longrightarrow+\infty}{\longrightarrow} 0$ for each $r \in \{D_{m^*}+1, \cdots, q\}$, we deduce from (4.11) that

$$\mathrm{FDR}(\hat{m}(K)) \underset{K\longrightarrow+\infty}{\longrightarrow} 0$$

As $P_r(K) \underset{K\longrightarrow+\infty}{\longrightarrow} 1$ for each $r \in \{D_{m^*}+1, \cdots, q\}$, we deduce that for all $C_1 \in]0,1[$, there exists $\tilde{L}_{C_1} > 0$ such that $\forall K > \tilde{L}_{C_1}$ and $\forall r \in \{D_{m^*}+1, \cdots, q\}$, we have $C_1 \leq P_r(K)$. For the following, we fix $C_1 \in]0,1[$.
Moreover, $P_r(K) \leq 1$ for each $r \in \{D_{m^*}+1, \cdots, q\}$. Hence, by using (4.31) and (4.36), we deduce that:

$$\forall K > \tilde{L}_{C_1}, \qquad \mathrm{FDR}(\hat{m}(K)) \geq C_1 \sum_{r=D_{m^*}+1}^{q}\left(\frac{r-D_{m^*}}{r}\underline{f}_r(K, \beta^*, \sigma^2)\right) \tag{4.41}$$

and

$$\forall K > 0, \qquad \mathrm{FDR}(\hat{m}(K)) \leq \sum_{r=D_{m^*}+1}^{q}\left(\frac{r-D_{m^*}}{r}\overline{f}_r(K, \beta^*, \sigma^2))\right). \tag{4.42}$$

- **Upper bound of $\overline{f}_r$:**
For each $r \in \{D_{m^*}+1, \cdots, q\}$ and for all $K > 0$:

$$\begin{aligned}
\overline{f}_r(K, \beta^*, \sigma^2)) &= 1 - \max\Bigg(\max_{\ell \in \{1,\cdots,r-D_{m^*}\}} \Big(F_{\chi^2(\ell)}(\ell K) \Big), \\
&\qquad \max_{\ell \in \{r-D_{m^*}+1,\cdots,r\}} \Big(F_{\chi^2(\ell)}\Big(\frac{\ell K}{2} - \sum_{k=r-\ell+1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2} \Big)\Big)\Bigg) \\
&= \min\Bigg(\min_{\ell \in \{1,\cdots,r-D_{m^*}\}} \Big(\mathbb{P}\big(X_\ell > \ell K\big)\Big), \\
&\qquad \min_{\ell \in \{r-D_{m^*}+1,\cdots,r\}} \Big(\mathbb{P}\big(Y_\ell > \frac{\ell K}{2} - \sum_{k=r-\ell+1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}\big)\Big)\Bigg) \\
&\qquad \text{with } X_\ell \sim \chi^2(\ell) \text{ and } Y_\ell \sim \chi^2(\ell) \\
&= \min\Bigg(\min_{\ell \in \{1,\cdots,r-D_{m^*}\}} \Big(\mathbb{P}\big(X_\ell - \ell > \ell K - \ell\big)\Big), \\
&\qquad \min_{\ell \in \{r-D_{m^*}+1,\cdots,r\}} \Big(\mathbb{P}\Big(Y_\ell - \ell > \frac{\ell(K-2)}{2} - \sum_{k=r-\ell+1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}\Big)\Big) \\
&\qquad \text{with } X_\ell \sim \chi^2(\ell) \text{ and } Y_\ell \sim \chi^2(\ell).
\end{aligned} \tag{4.43}$$

So, for each $r \in \{D_{m^*}+1, \cdots, q\}$ and for all $K > 0$:

$$\overline{f}_r(K, \beta^*, \sigma^2)) \leq \min_{\ell \in \{1,\cdots,r-D_{m^*}\}} \Big(\mathbb{P}\big(X_\ell - \ell > \ell K - \ell\big)\Big), \qquad \text{with } X_\ell \sim \chi^2(\ell)$$

We recall the exponential inequality of [Laurent and Massart, 2000] for chi-squared distributions of mean $\ell \in \mathbb{N}^*$:

$$\text{For any positive } x \text{ and for } X \sim \chi^2(\ell), \qquad \mathbb{P}\Big(X - \ell > 2\sqrt{\ell x} + 2x\Big) \leq e^{-x} \tag{4.44}$$

We apply (4.44) for each $\ell = 1, \cdots, (r - D_{m^*})$ with $x = \frac{\ell}{4}\Big(1 - \sqrt{2K-1}\Big)^2$ which is one solution of $2\sqrt{\ell x} + 2x = \ell K - \ell$ when $K > 1$. We obtain for all $K > 1$:

$$\begin{aligned}
\min_{\ell \in \{1,\cdots,r-D_{m^*}\}} \Big(\mathbb{P}\big(X_\ell - \ell > \ell K - \ell\big)\Big) &\leq \min_{\ell=1,\cdots,(r-D_{m^*})} \Big(e^{-\frac{\ell}{4}\left(1-\sqrt{2K-1}\right)^2} \Big) \\
&\leq e^{\frac{(r-D_{m^*})\sqrt{2K-1}}{2}} \min_{\ell=1,\cdots,(r-D_{m^*})} \Big(e^{-\frac{\ell K}{2}} \Big) \\
&= e^{\frac{(r-D_{m^*})\sqrt{2K-1}}{2}} e^{-\frac{(r-D_{m^*})K}{2}}.
\end{aligned} \tag{4.45}$$

So, from (4.42) and (4.45), we obtain for each $r \in \{D_{m^*}+1, \cdots, q\}$ and for all $K > 1$:

$$\begin{aligned}\mathrm{FDR}(\hat{m}(K)) &\leq \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} e^{\frac{(r-D_{m^*})\sqrt{2K-1}}{2}} e^{-\frac{(r-D_{m^*})K}{2}} \right) \\ &\leq e^{-\frac{K}{2}} \sum_{r=D_{m^*}+1}^{q} \left(\frac{r - D_{m^*}}{r} e^{\frac{(r-D_{m^*})\sqrt{2K-1}}{2}} \right) \end{aligned} \tag{4.46}$$

For all $\eta > 0$ and $r \in \{D_{m^*}+1, \cdots, q\}$, $e^{\frac{(r-D_{m^*})\sqrt{2K-1}}{2}} = \underset{K \longrightarrow +\infty}{o}\left(e^{\eta K}\right)$. Hence, (4.46) is $\underset{K \longrightarrow +\infty}{o}\left(e^{-K(\frac{1}{2}-\eta)}\right), \forall \eta > 0$.
Hence, $\forall \eta > 0$

$$\mathrm{FDR}(\hat{m}(K)) = \underset{K \longrightarrow +\infty}{o}\left(e^{-K(\frac{1}{2}-\eta)}\right),$$

which allows to obtain (4.17).

- - **Proof of Remark 4.2**:
The inequalities (4.41) and (4.42) are also true when $K \longrightarrow +\infty$ and $\sigma \longrightarrow 0$ with $\frac{1}{\sigma} = \underset{\sigma \longrightarrow 0}{o}(\sqrt{K})$. To obtain the finest asymptotic upper bound (4.20), we start from the equation (4.43) and we consider the second term. In the same way as previously, we apply (4.44) for each $\ell = r - D_{m^*} + 1, \cdots, r$ with $x = \frac{\ell}{4}\left(1 - \sqrt{K - 1 - \frac{2}{\ell} \sum_{k=r-\ell+1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}}\right)^2$ which is one solution of $2\sqrt{\ell x} + 2x = \frac{\ell(K-2)}{2} - \sum_{k=r-\ell+1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}$ when $\sigma^2(K-1) > \frac{2}{r-D_{m^*}+1} \sum_{k=1}^{D_{m^*}} \langle X\beta^*, u_k\rangle^2 + 2$. This condition is valid since $\sigma \longrightarrow 0$ with $\frac{1}{\sigma} = \underset{\sigma \longrightarrow 0}{o}(\sqrt{K})$ leading to $\frac{1}{\sigma^2} = \underset{\sigma \longrightarrow 0}{o}(K)$ and so $\sigma^2(K-1) \longrightarrow +\infty$ when $K \longrightarrow +\infty$. We obtain for all $K > 0$ such that $\sigma^2(K-1) > \frac{2}{r-D_{m^*}+1} \sum_{k=1}^{D_{m^*}} \langle X\beta^*, u_k\rangle^2 + 2$:

$$\begin{aligned} &\min_{\ell \in \{r-D_{m^*}+1, \cdots, r\}} \left(\mathbb{P}\big(Y_\ell - \ell > \frac{\ell(K-2)}{2} - \sum_{k=r-\ell+1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}\big)\right) \\ &\leq \min_{\ell \in \{r-D_{m^*}+1, \cdots, r\}} \left(e^{-\frac{\ell}{4}\left(1 - \sqrt{K-1-\frac{2}{\ell}\sum_{k=r-\ell+1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}}\right)^2} \right) \\ &\underset{(*)}{\leq} e^{\frac{1}{2}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}} e^{\frac{r}{2}\sqrt{K-1-\frac{2}{r}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}}} \min_{\ell = r-D_{m^*}+1, \cdots, r} \left(e^{-\frac{\ell K}{4}} \right) \\ &= e^{\frac{1}{2}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}} e^{\frac{r}{2}\sqrt{K-1-\frac{2}{r}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}}} e^{-\frac{rK}{4}}. \end{aligned} \tag{4.47}$$

(*) come from the fact that a minimum into a set is smaller than any value in the set. We choose the value corresponding for $\ell = 0$.
So, from (4.42), (4.45) and (4.47), we obtain for each $r \in \{D_{m^*}+1, \cdots, q\}$ and for all $K > 1$ respecting $\sigma^2(K-1) > \frac{2}{r-D_{m^*}+1}\sum_{k=1}^{D_{m^*}} \langle X\beta^*, u_k\rangle^2 + 2$:

$$\begin{aligned}
\mathrm{FDR}(\hat{m}(K)) \leq & \sum_{r=D_{m^*}+1}^{q} \Bigg(\frac{r-D_{m^*}}{r} \min \Bigg(e^{\frac{(r-D_{m^*})\sqrt{2K-1}}{2}} e^{-\frac{(r-D_{m^*})K}{2}}, \\
& e^{\frac{1}{2}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}} e^{\frac{r}{2}\sqrt{K-1-\frac{2}{r}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}}} e^{-\frac{rK}{4}} \Bigg) \Bigg) \\
= & \min \Bigg(\sum_{r=D_{m^*}+1}^{q} \Bigg(\frac{r-D_{m^*}}{r} e^{\frac{(r-D_{m^*})\sqrt{2K-1}}{2}} e^{-\frac{(r-D_{m^*})K}{2}} \Bigg), \\
& \sum_{r=D_{m^*}+1}^{q} \Bigg(\frac{r-D_{m^*}}{r} e^{\frac{1}{2}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}} e^{\frac{r}{2}\sqrt{K-1-\frac{2}{r}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}}} e^{-\frac{rK}{4}} \Bigg) \Bigg) \\
\leq & \min \Bigg(e^{-\frac{K}{2}} \sum_{r=D_{m^*}+1}^{q} \Bigg(\frac{r-D_{m^*}}{r} e^{\frac{(r-D_{m^*})\sqrt{2K-1}}{2}} \Bigg), \\
& e^{-\left(\frac{(D_{m^*}+1)K}{4} - \frac{1}{2\sigma^2}\sum_{k=1}^{D_{m^*}} \langle X\beta^*, u_k\rangle^2\right)} \sum_{r=D_{m^*}+1}^{q} \Bigg(\frac{r-D_{m^*}}{r} e^{\frac{r}{2}\sqrt{K-1-\frac{2}{r}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}}} \Bigg) \Bigg)
\end{aligned} \tag{4.48}$$

For all $\eta > 0$ and $r \in \{D_{m^*}+1, \cdots, q\}$, $e^{\frac{(r-D_{m^*})\sqrt{2K-1}}{2}} = \underset{K \longrightarrow +\infty}{o}\left(e^{\eta K}\right)$, independently of the value of σ^2. Hence, the first term in (4.48) is $o\left(e^{-K(\frac{1}{2}-\eta)}\right), \forall \eta > 0$ when $K \longrightarrow +\infty$ and $\sigma \longrightarrow 0$ with $\frac{1}{\sigma} = \underset{\sigma \longrightarrow 0}{o}(\sqrt{K})$.

For all $r \in \{D_{m^*}+1, \cdots, q\}$, $e^{\frac{r}{2}\sqrt{K-1-\frac{2}{r}\sum_{k=1}^{D_{m^*}} \frac{\langle X\beta^*, u_k\rangle^2}{\sigma^2}}} \leq e^{\frac{r}{2}\sqrt{K}}$. Moreover, for all $\tilde{\eta} > 0$ and $r \in \{D_{m^*}+1, \cdots, q\}$, $e^{\frac{r}{2}\sqrt{K}} = \underset{K \longrightarrow +\infty}{o}\left(e^{\tilde{\eta} K}\right)$, independently of the value of σ^2. Hence, the second term in (4.48) is $o\left(e^{-\left(K\frac{(D_{m^*}+1-\tilde{\eta})}{4} - \frac{1}{2\sigma^2}\sum_{k=1}^{D_{m^*}} \langle X\beta^*, u_k\rangle^2\right)}\right)$, $\forall \tilde{\eta} > 0$ when $K \longrightarrow +\infty$ and $\sigma \longrightarrow 0$ with $\frac{1}{\sigma} = \underset{\sigma \longrightarrow 0}{o}(\sqrt{K})$.

Hence,

$$\mathrm{FDR}(\hat{m}(K)) \leq \min\left(o\left(e^{-K(\frac{1}{2}-\eta)}\right), o\left(e^{-\left(K\frac{(D_{m^*}+1-\tilde{\eta})}{4}-\frac{1}{2\sigma^2}\sum_{k=1}^{D_{m^*}}\langle X\beta^*,u_k\rangle^2\right)}\right)\right)$$
$$= o\left(e^{-\left(K\frac{(D_{m^*}+1-\tilde{\eta})}{4}-\frac{1}{2\sigma^2}\sum_{k=1}^{D_{m^*}}\langle X\beta^*,u_k\rangle^2\right)}\right)$$

$\forall(\eta,\tilde{\eta})>0$ when $K\longrightarrow+\infty$ and $\sigma\longrightarrow 0$ with $\frac{1}{\sigma}=\underset{\sigma\longrightarrow 0}{o}(\sqrt{K})$; which allows us to obtain (4.20).

- Lower bound of f_r:
From (4.40) and (4.41), we obtain:

$$\forall K>\tilde{L}_{C_1},\quad \mathrm{FDR}(\hat{m}(K)) \geq C_1\sum_{r=D_{m^*}+1}^{q}\left(\frac{r-D_{m^*}}{r}\frac{2\sqrt{2}}{\sqrt{\pi}\left(\sqrt{rK}+\sqrt{rK+4}\right)}e^{-\frac{rK}{2}}\right)$$
$$\geq C_1\frac{2\sqrt{2}}{\sqrt{\pi}\left(\sqrt{qK}+\sqrt{qK+4}\right)}\frac{1}{D_{m^*}+1}\sum_{r=D_{m^*}+1}^{q}\left(e^{-\frac{rK}{2}}\right)$$
$$\underset{(***)}{\geq} C_1\frac{2\sqrt{2}}{\sqrt{\pi}\left(\sqrt{qK}+\sqrt{qK+4}\right)}\frac{1}{D_{m^*}+1}e^{-\frac{(D_{m^*}+1)K}{2}}$$
$$= \frac{2\sqrt{2}C_1}{\sqrt{\pi}(D_{m^*}+1)}\frac{1}{\sqrt{qK}+\sqrt{qK+4}}e^{-K\frac{(D_{m^*}+1)}{2}}$$

(***) is true since each term in the sum is positive, so, the sum is larger than one of them.
For all $\eta>0$, $\exists\tilde{C}_\eta>0$, $\exists\tilde{L}_\eta>0$ such that $\forall K>\tilde{L}_\eta$, we have $\tilde{C}_\eta e^{-\eta K}\leq\frac{1}{\sqrt{qK}+\sqrt{qK+4}}$.
So,

$$\forall\eta>0,\ \exists\tilde{C}_\eta>0,\ \exists\tilde{L}_\eta>0,\ \forall K>\max\left(\tilde{L}_{C_1},\tilde{L}_\eta\right),$$
$$\mathrm{FDR}(\hat{m}(K)) \geq \frac{2\sqrt{2}C_1}{\sqrt{\pi}(D_{m^*}+1)}\tilde{C}_\eta e^{-K\left(\frac{D_{m^*}+1+2\eta}{2}\right)}$$

which gives (4.18) with $C_\eta=\frac{2\sqrt{2}C_1}{\sqrt{\pi}(D_{m^*}+1)}\tilde{C}_\eta$ and $L_\eta=\max\left(\tilde{L}_{C_1},\tilde{L}_\eta\right)$.

Formula (4.19) automatically follows from (4.17) and (4.18). □

4.5.5 General bounds.

Proof of Corollary 4.3.
By taking $u_j=X_j,\ \forall j\in\{1,\cdots,q\}$, then $(X_1,\cdots,X_q,u_{q+1},\cdots,u_n)$ is an orthonormal basis

of $\mathbb{R}^n$. Consequently, $\forall j \in \{1, \cdots, q\}$, $\langle X\beta^*, u_j \rangle = \langle X\beta^*, X_j \rangle = \beta_j^*$, which concludes the proof. $\square$

4.6 Conclusions.

The variable selection procedure in a high-dimensional Gaussian linear regression with sparsity assumption is commonly used either to identify a set of variables with prediction performances or to avoid the selection of non active variables. The first goal is processed by controlling the PR via the penalized least-squared procedures; the second one is processed by controlling the FDR via the multiple testing methods. On the one hand, controlling the PR tends to select too many variables, including non active ones; on the other hand, controlling the FDR tends to select too few variables, leaving out some active ones. We propose to work with both PR and FDR cost functions simultaneously. The goal is to get prediction performances with a low number of non active variables.
In this way, we propose to calibrate the hyperparameter K of the penalty function (4.4) in the model selection procedure with a known variance σ^2 and ordered variable selection assumptions. Firstly, the FDR function's non-asymptotic lower and upper bounds with respect to $K > 0$ are obtained. These bounds are easily implementable and do not depend on the data but only on the model parameters. The studied asymptotic behavior suggests that bounds are optimal. Secondly, these theoretical results are used for the applications. An extensive simulation study justifies the estimation of σ^2 by the slope heuristic methods; and shows that $\hat{\beta}_{\hat{m}(4)}$ can reasonably replace β^* in the FDR bounds. Thirdly, we propose an algorithm to calibrate the hyperparameter K in the penalty function. It is based on completely data-driven terms: the estimated risk and the estimated upper bound $\tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ of the FDR. The hyperparameter K is calibrated from the dataset as the smallest constant $K \geq 2$ belonging to the intersection of the interval of K with the smallest values of estimated risk and the interval of K where $\tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ values are strictly positive and smaller than a given threshold. Our large simulations study valid our data-driven penalty function since the calibrated hyperparameter $K \geq 2$ leads to a selected model ensuring a low value of both theoretical PR and FDR criteria providing good prediction performance properties and containing a low proportion of non active variables. The calibrated hyperparameter K is between 2.7 and 4.8, so strictly larger than the commonly used constant $K = 2$. The final selected variables subset corresponds to similar predictive performances as for $K = 2$ but with a FDR value possibly lower than 0.05. So, this method can get at least comparable results to the multiple testing procedures where the unique aim is to control the FDR quantity and the common threshold is 0.05.

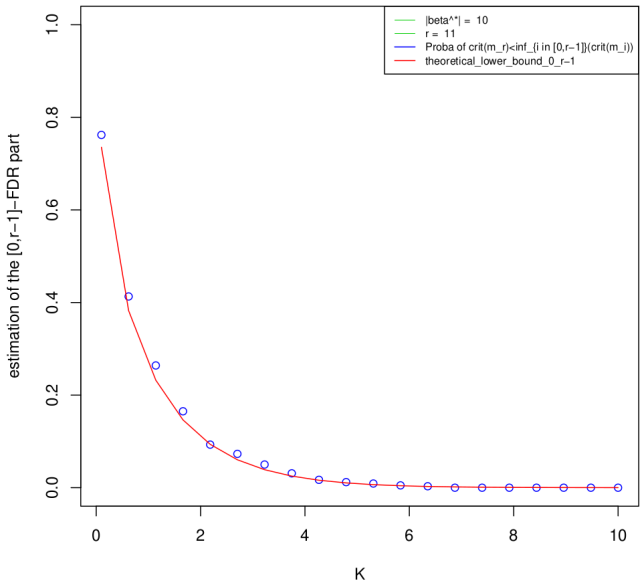
Our algorithm is based on theoretical control of both PR and FDR but nothing says that our choices of complete data-driven terms, the estimated risk and $\tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ terms, to estimate theoretical ones are the best ones. Moreover, the hyperparameter K calibration is not necessarily the best either. So, a perspective is the construction of other algorithms based on the theoretical results. It can lead to other data-driven terms and other calibrations of K, to obtain, for instance, some properties on K (stability, robustness,...).

Some limitations of our results are the following. If $D_{\hat{m}(K)} = q$ for one $K > 1$, our lower and upper bounds are estimated to 0 meaning that if $D_{\hat{m}(K)} = q$, which can frequently happen for small K values, a distinction between $D_{m^*} = q$ and $D_{m^*} < q$ is not possible without additional arguments. Moreover, a perspective is to relax some assumptions. Indeed, m^* is unknown in practice and variables do not come with an order in the sense of (4.1). Future work will generalize our results from the ordered variable selection procedure to complete variable selection one without assumption on m^*. Our work can also be generalized to random collection or a non-fixed design. These extensions are much more intricate.
A possible opening is to study the potential properties of the hyperparameter K provided by our data-driven hyperparameter calibration. A constant K given a low value of both PR and FDR criteria in the model selection could respect some theoretical characteristics. Intuitively, the larger $|\beta^*|$ is, the smaller the penalty should be and the fewer non active variables are selected, the smaller the FDR is. So, the larger $|\beta^*|$ is, the smaller the multiplicative constant K should be. Hence, the hyperparameter K satisfying a trade-off between PR and FDR could theoretically be expressed as a function of $|\beta^*|$. Finally, another possible extension is to study the power of the procedure by adding a study of the FNR (false negative rate) with respect to $K > 0$ in our process. Indeed, the FDR gives only partial information about the quality of the selected variables set since the focus is only on the selected variables. Conversely, the FNR focuses only on the non-selected variables. Optimal control of the three criteria PR, FDR and FNR should not be hoped (see [Su et al., 2017]) but ensuring a reasonable triple control of these criteria can provide better performances and a more stable selected variables set.

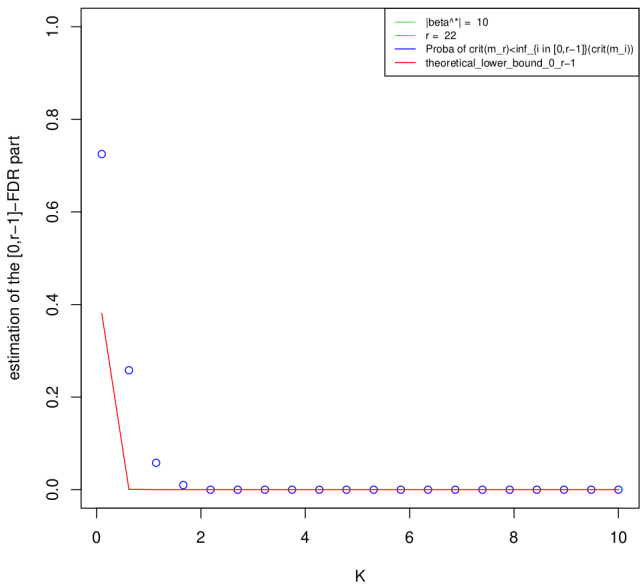
4.7 Appendix: More detailed study of Theorem 4.1 in the orthogonal case of Corollary 4.3.

In Subsection 4.3.3, we observe that the curves of functions $\mathrm{FDR}\big(\hat{m}(K)\big), b(K,\beta^*,\sigma^2)$ and $B(K,\beta^*,\sigma^2)$ for all $K>0$ are almost overlapped. Moreover, the asymptotic behaviors (4.17) and (4.18) suggest that the obtained convergence rate is optimal. To further support this intuition, we evaluate the inequalities performed in Theorem 4.1 to obtain the $b(K,\beta^*,\sigma^2)$ and $B(K,\beta^*,\sigma^2)$ bounds. This corresponds to an evaluation of the losses incurred when we use respectively $\underline{f}_r$ and $\overline{f}_r$ instead of the Q_r quantity. Let us recall that for all $K>0$, $Q_r(K,\beta^*,\sigma^2)$ is equal to $\mathbb{P}\left(\left\{\bigcap_{\ell=0}^{r-1}\{\mathrm{crit}_K(m_r)<\mathrm{crit}_K(m_\ell)\}\right\}\right)$ and $P_r(K)$ is equal to $\mathbb{P}\left(\left\{\bigcap_{\ell=r+1}^{q}\{\mathrm{crit}_K(m_r)<\mathrm{crit}_K(m_\ell)\}\right\}\right)$. We estimate empirically this two quantities on our simulated datasets by counting respectively the number of times that $\mathrm{crit}_K(m_r)<\mathrm{crit}_K(m_\ell), \forall l\in\{0,r-1\}$ over the 1000 iterations and the number of times that $\mathrm{crit}_K(m_r)<\mathrm{crit}_K(m_\ell), \forall \ell\in\{r+1,q\}$ over the 1000 iterations. Figures 4.11 and 4.12 compare respectively the functions $\underline{f}_r(K,\beta^*,\sigma^2)$ and $\overline{f}_r(K,\beta^*,\sigma^2))$ with the empirical estimation of $Q_r(K,\beta^*,\sigma^2)$ for different values of $r\in\{D_{m^*}+1,\cdots,q\}$ and for all $K>0$. Concerning the $P_r(K)$ quantity, its formula (4.12) does not depend on the data as soon as r is given. For each $r>0$, we estimate the quantity $\mathbb{P}\left(\bigcap_{\ell=r+1}^{q}\left\{\sum_{k=r+1}^{\ell}Z_k^2<K(\ell-r)\right\}\right)$ by generated 5000 independent standard Gaussian vectors $\Big(Z_k\Big)_{k\in\{r+1,\ldots,q\}}$ and by counting for each vector the number of time that $Z_k^2<K(\ell-r)$ for each $\ell\in\{r+1,\cdots,q\}$. Figure 4.13 compares the empirical estimation function of $P_r(K)$ with its formula (4.12) estimation for different values of $r\in\{D_{m^*}+1,\cdots,q\}$ and for all $K>0$. We observe that for all $K>0$, curves are all almost overlapping, even more so when r is small and especially for $K\geq 2$. These observations suggest that the used inequalities are not brutal. We propose in supplementary material available in [8] a complementary study of Q_r and P_r with different values of σ^2 ($\sigma^2=1, 0.1$ and 4) since for a fixed r, theirs bounds and their empirical estimations depend on σ^2.

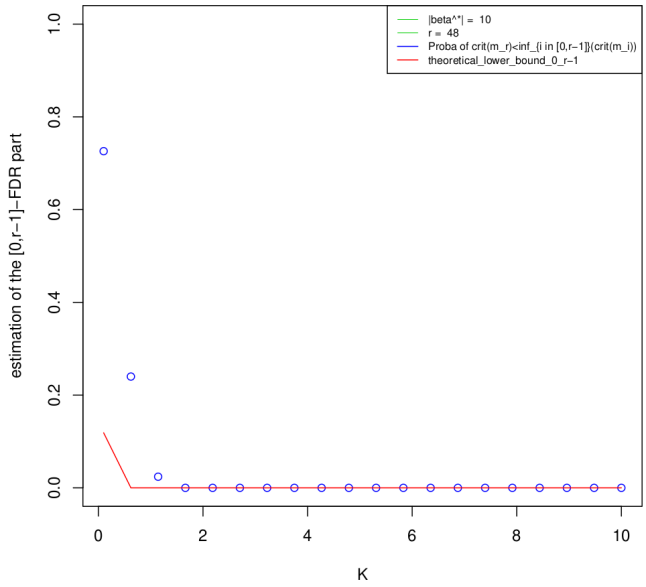
[8] `https://sites.google.com/view/placroix/research`



r=11

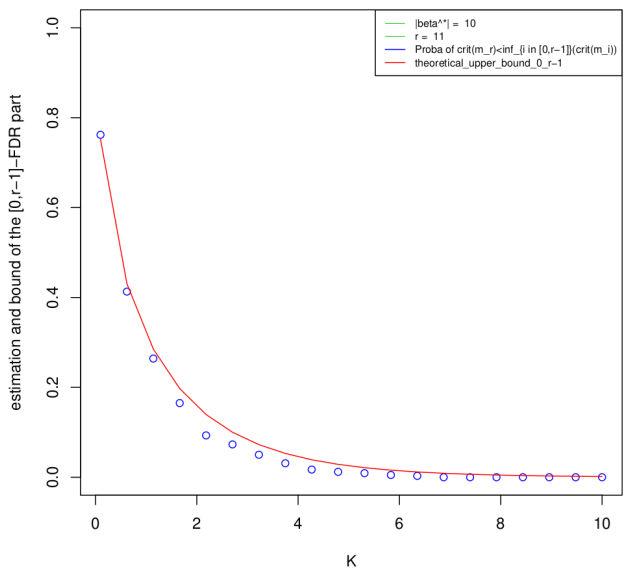


r=22

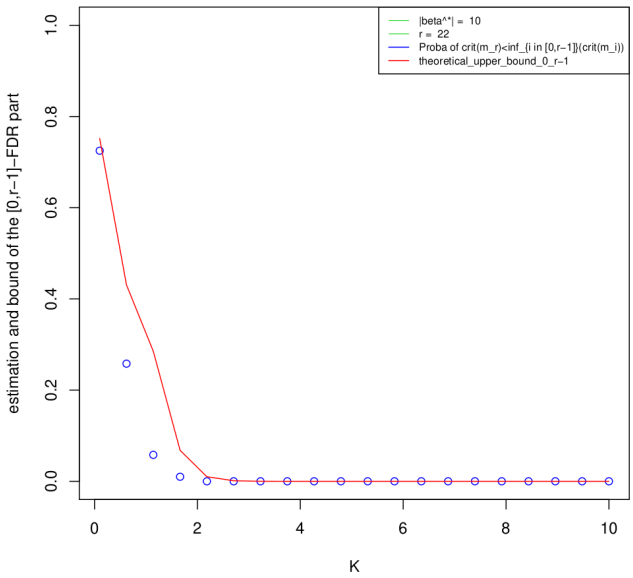


r=48

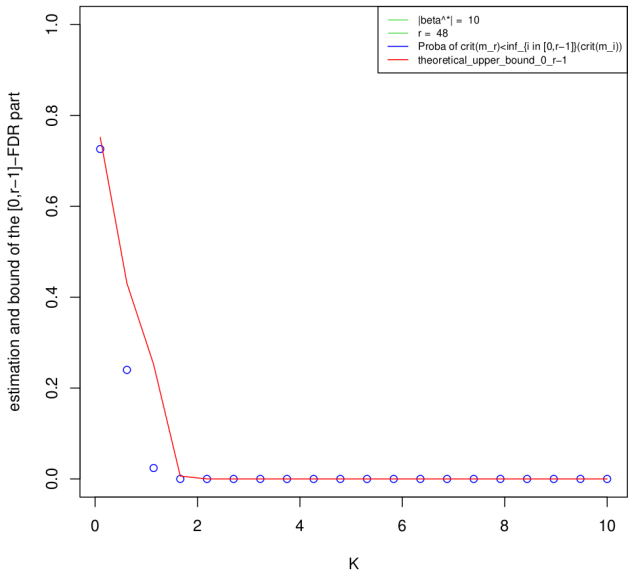
Figure 4.11: Comparison for all $K > 0$ of the function $\underline{f}_r(K, \beta^*, \sigma^2)$ with the empirical estimator of $\mathbb{P}\left(\bigcap_{\ell=0}^{r-1}\left\{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\right\}\right)$ when $\sigma^2 = 1$.



r=11

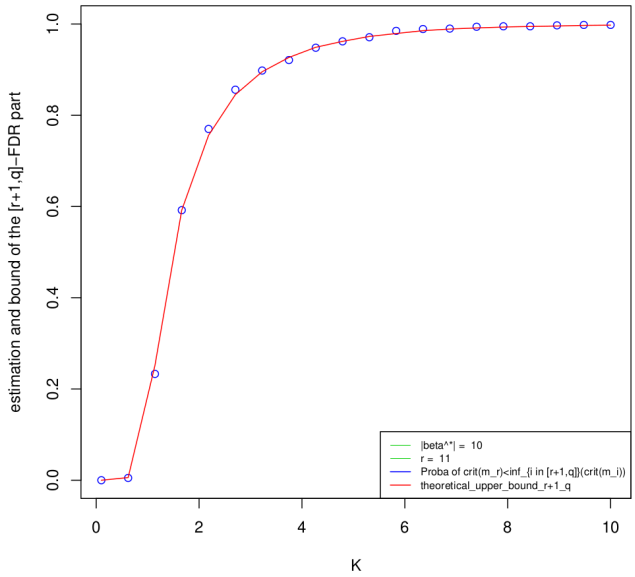


r=22

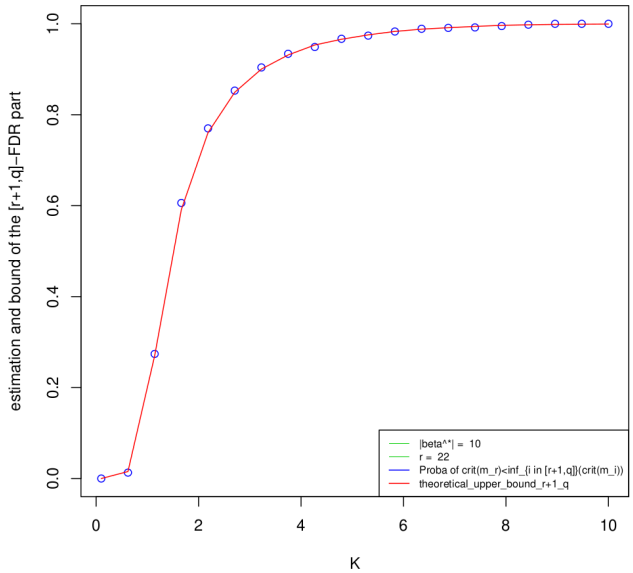


r=48

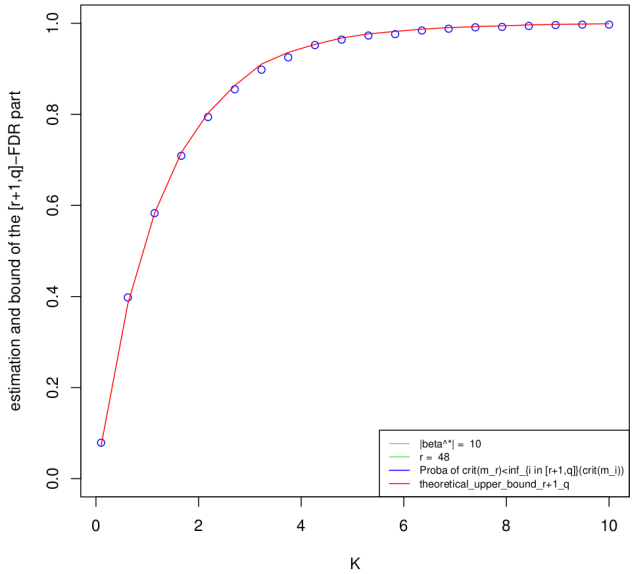
Figure 4.12: Comparison for all $K > 0$ of the function $\overline{f}_r(K, \beta^*, \sigma^2)$ with the empirical estimator of $\mathbb{P}\left(\bigcap_{\ell=0}^{r-1}\left\{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\right\}\right)$ when $\sigma^2 = 1$.



r=11



r=22



r=48

Figure 4.13: Comparison for all $K > 0$ of the $P_r(K)$ formula (4.12) with the empirical estimator of $\mathbb{P}\left(\bigcap_{\ell=r+1}^{q}\left\{\mathrm{crit}_K(m_r) < \mathrm{crit}_K(m_\ell)\right\}\right)$ when $\sigma^2 = 1$.

4.8 Appendix: Variation of some parameters in the orthogonal case of Corollary 4.3.

Impacts of parameters variation. When we focus on each scenario, especially concerning the $\tilde{B}(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ terms and the result of our proposed data-driven hyperparameter calibration, the higher the D_{m^*} value is, the smaller the empirical FDR is but the larger the MSE for large K is. Moreover, the relative change functions decreases when D_m^* increases, as well as the relative standard deviation ones which remain smaller than 0.5. This can be explained since the higher D_{m^*}, the smaller the number of non active variables, and so the number of the selected non active variables. This leads to a smaller FDR. In the opposite trend, the MSE increases with D_{m^*} since penalization tends to select too few variables, especially even K moves away from 2, leading to less accurate prediction performances. As expected, concerning the scenario (ii), when coefficients are smaller than the amplitude of the noise (the second configuration), values of the relative change for the $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ bounds explode (until 10^5); and the relative standard deviation values increase until exceed 1. The best results are obtained for the first β^* configuration of the scenario (ii), but with the third one, results still remain reasonable. The relative standard deviation functions increase after $\tilde{K} \geq 4$ whereas in all other scenarios, functions always decrease when $\tilde{K}$ increases. As for the scenario (iii), we observe unsurprisingly that the higher the value of n, the smaller the relative change, the smaller the relative standard deviation, and the tighter the confidence interval of PR (< 0.04 for $n = 30$). However, we note that the computational time to estimate the bounds was significantly higher for $n = 300$. Lastly, concerning the scenario (iv), as expected, the higher the noise amplitude, the larger the confidence interval for the PR (< 0.45 for $\sigma^2 = 4$), the higher the relative change (which equals 0 when $\sigma^2 = 0.1$ and around 2 when $\sigma^2 = 4$) and the higher the relative standard deviation. However, values remain reasonable even for $\sigma^2 = 4$ excepted for the MSE values which are always larger than 5.

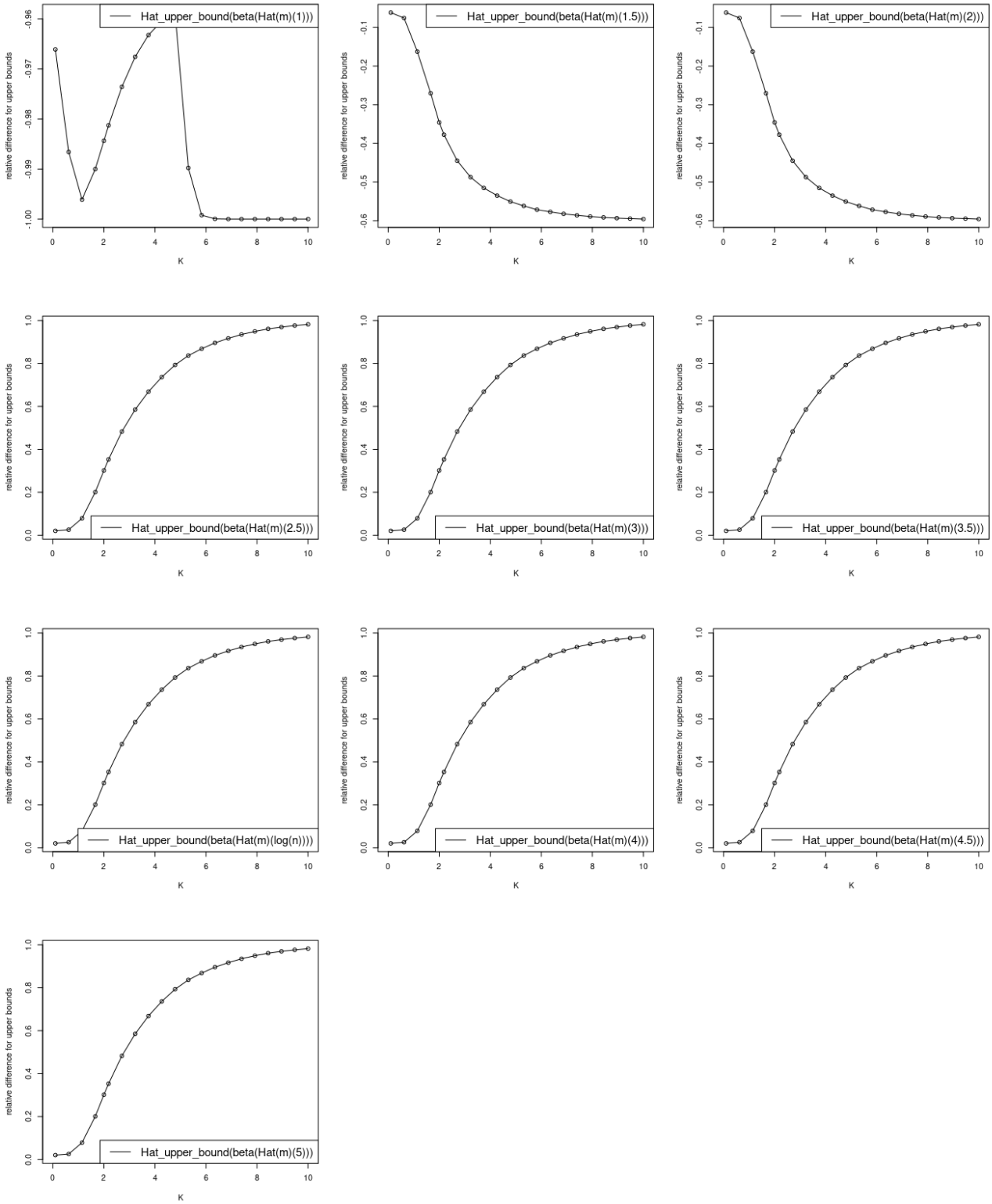


Figure 4.14: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from only one dataset. These plots are obtained with the toy data set described in Subsection 4.3.1 for $|\beta^*| = 1$.

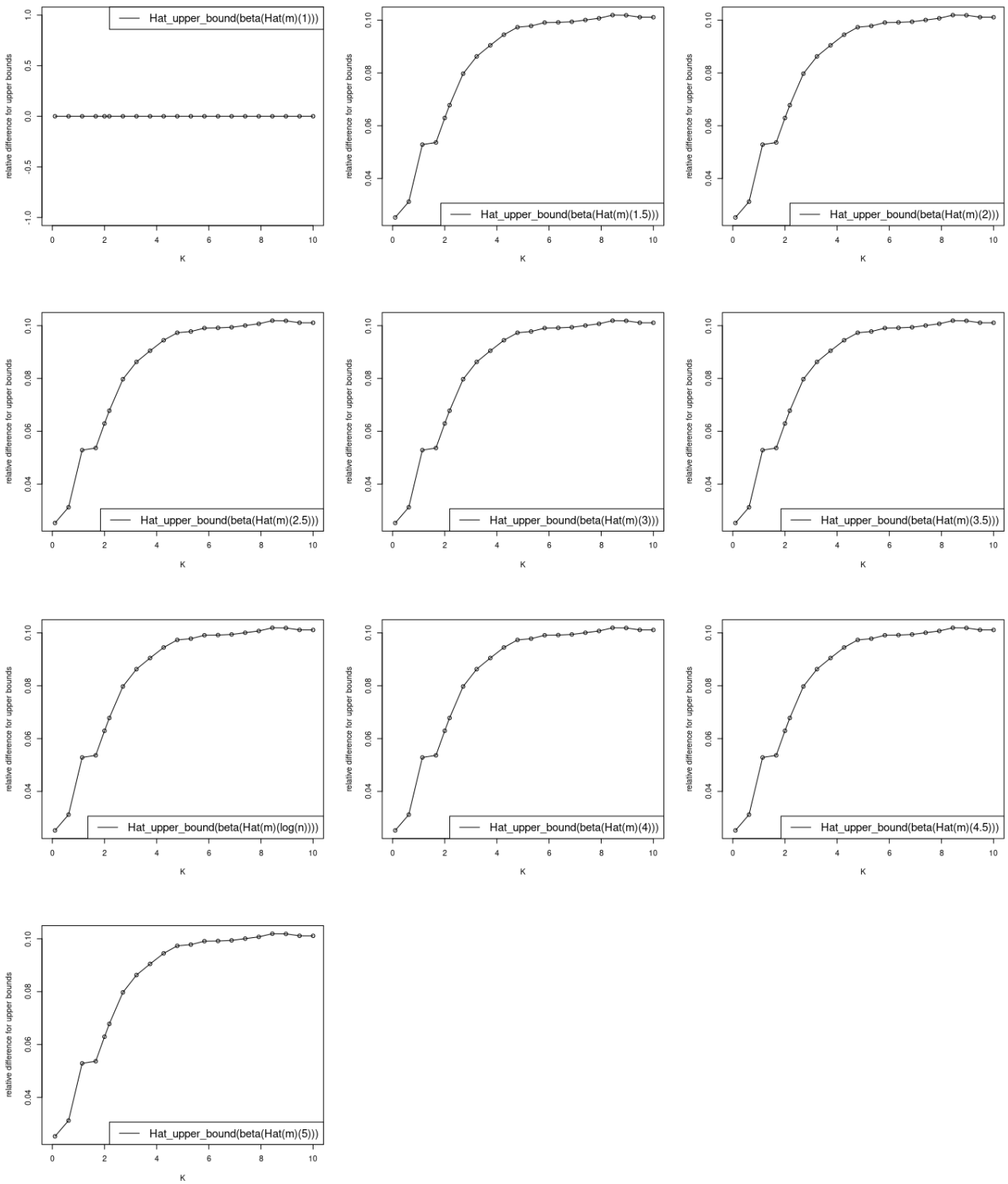


Figure 4.15: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from only one dataset. These plots are obtained with the toy data set described in Subsection 4.3.1 for $|\beta^*| = 10$.

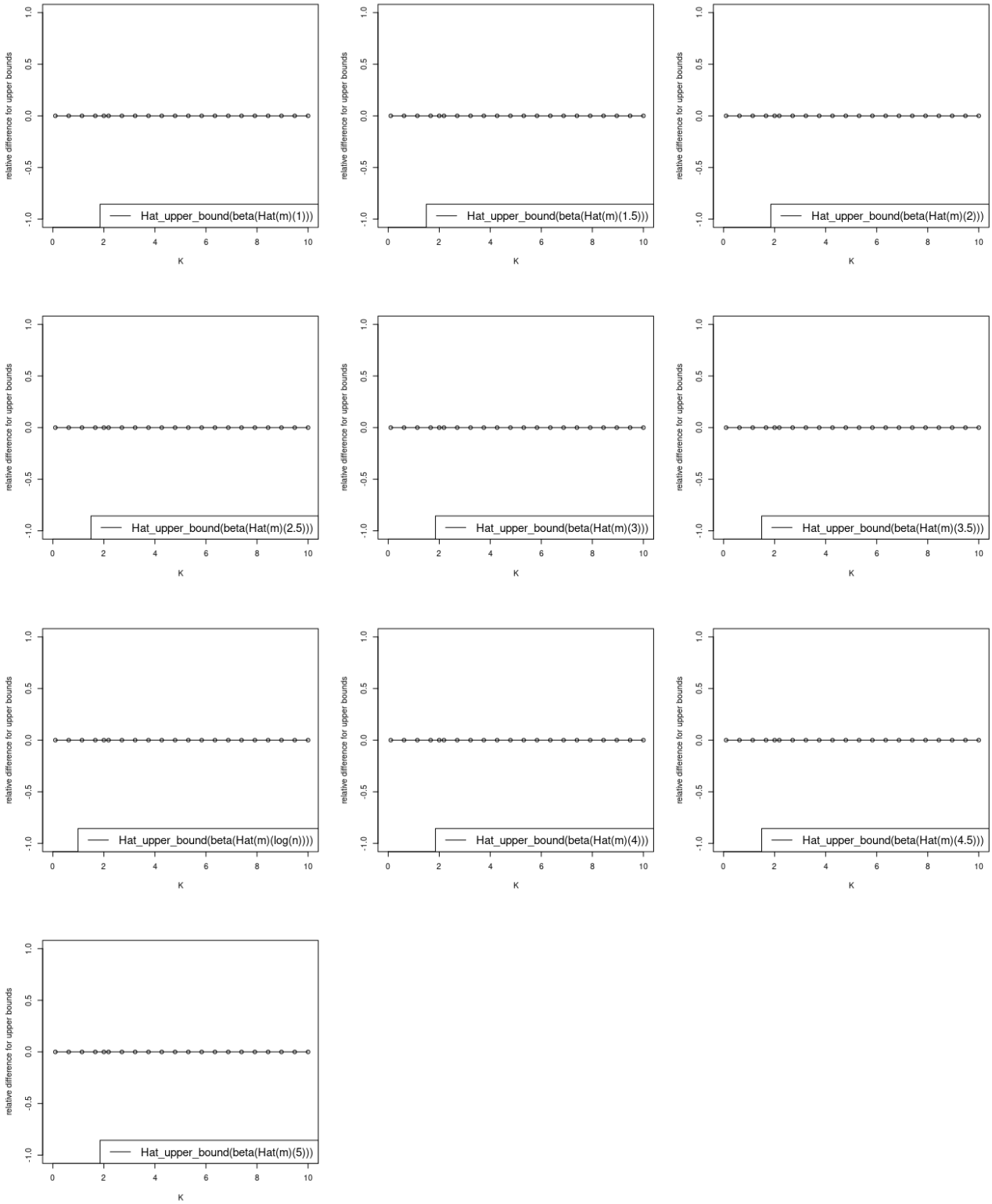


Figure 4.16: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from only one dataset. These plots are obtained with the toy data set described in Subsection 4.3.1 for $|\beta^*| = 20$.

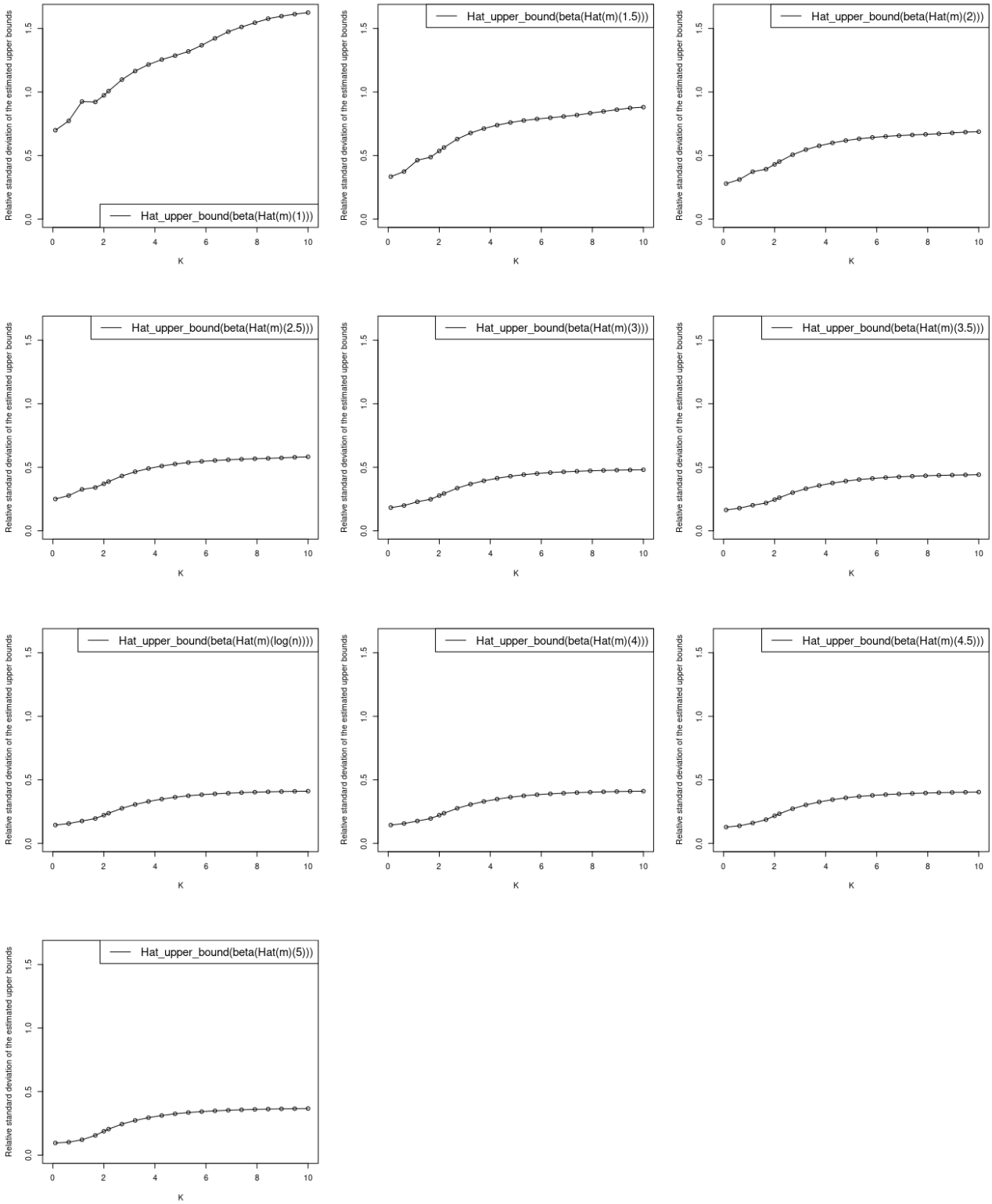


Figure 4.17: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. These plots are obtained with the toy data set described in Subsection 4.3.1 for $|\beta^*| = 1$.

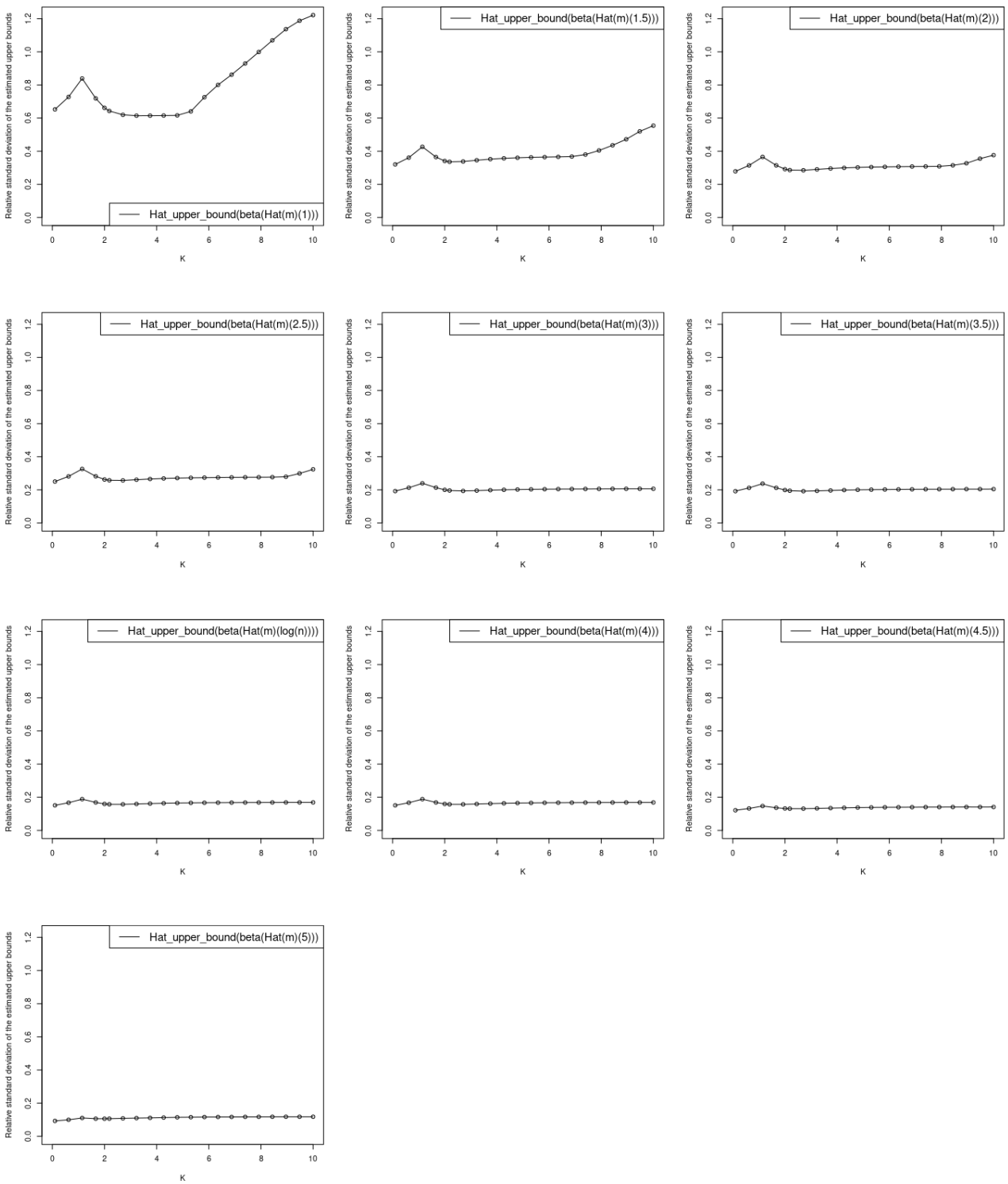


Figure 4.18: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. These plots are obtained with the toy data set described in Subsection 4.3.1 for $|\beta^*| = 10$.

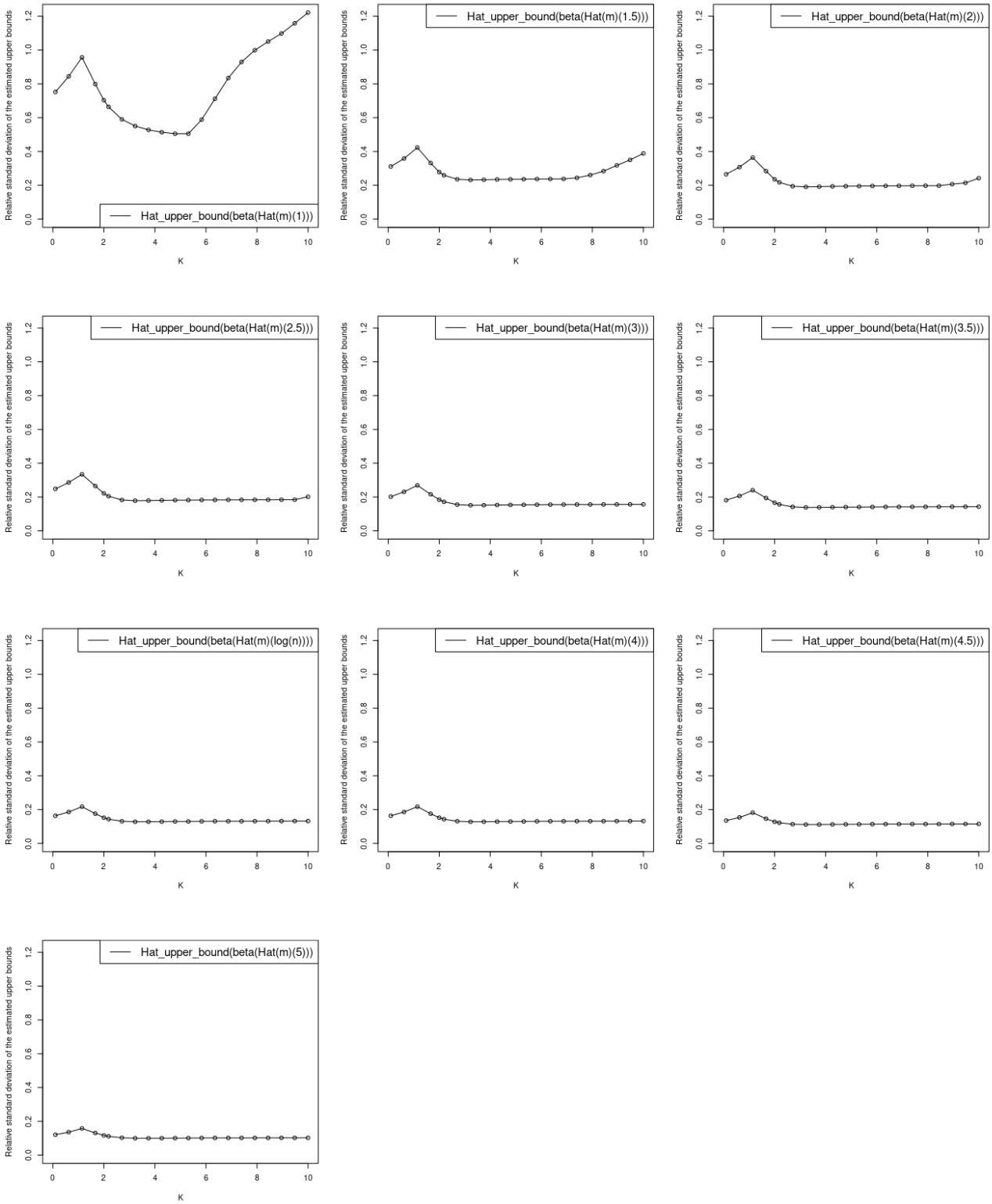
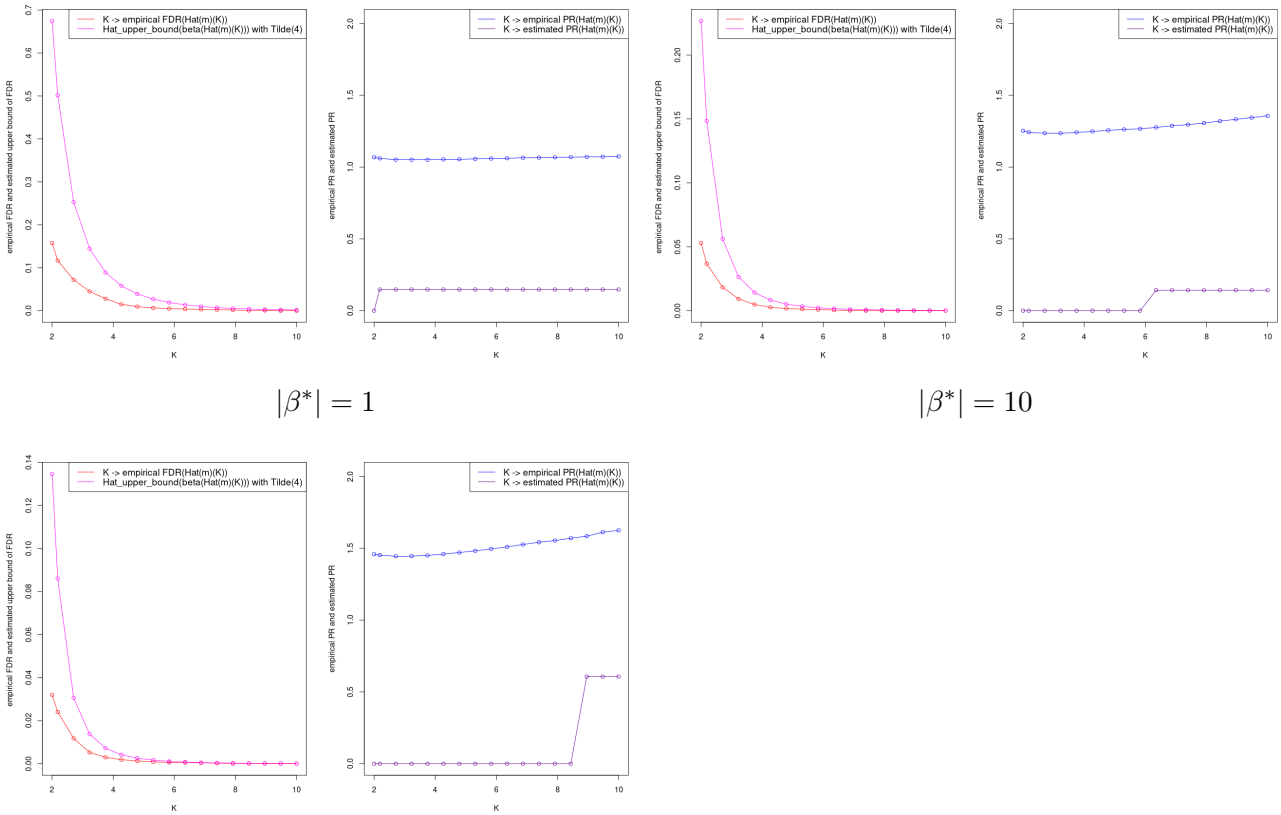


Figure 4.19: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. These plots are obtained with the toy data set described in Subsection 4.3.1 for $|\beta^*| = 20$.



$|\beta^*| = 1$ $|\beta^*| = 10$

$|\beta^*| = 20$

Figure 4.20: Curves of the empirical estimation functions of $\mathrm{FDR}\big(\hat{m}(K)\big)$ and $\mathrm{PR}\big(\hat{m}(K)\big)$ for all $K > 0$ by using 1000 datasets and curves of the estimated risk (4.22) and the $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ function obtained in Corollary 4.3 by replacing β^* by $\hat{\beta}_{\hat{m}(4)}$. These two last plots are obtained from only one dataset. These plots are obtained with the toy data set described in Subsection 4.3.1 for $|\beta^*| = 1, 10, 20$.

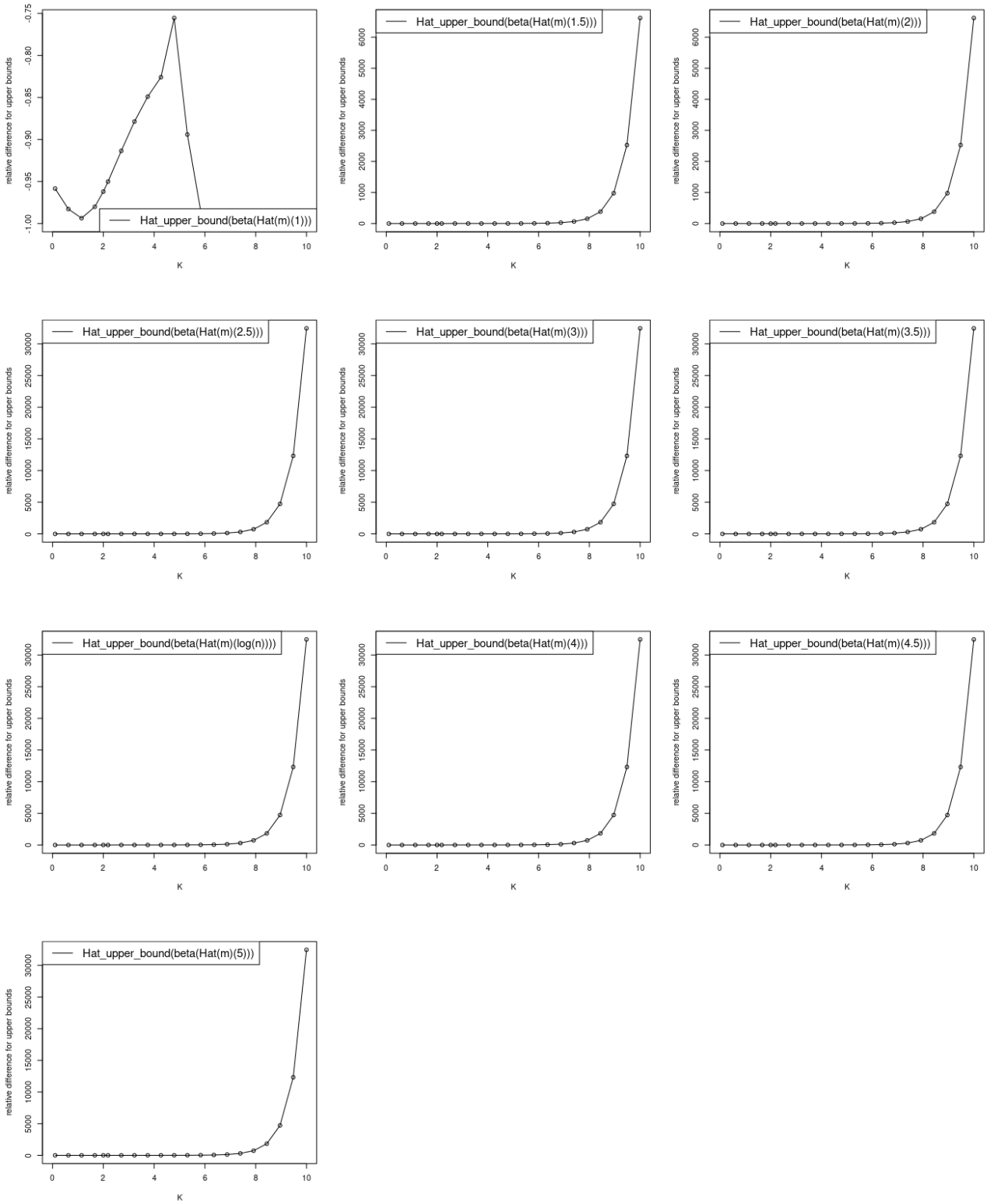


Figure 4.21: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from only one dataset. These plots are obtained with the toy data set described in Subsection 4.3.1 for $\beta^*_{10} = \frac{2}{10}$.

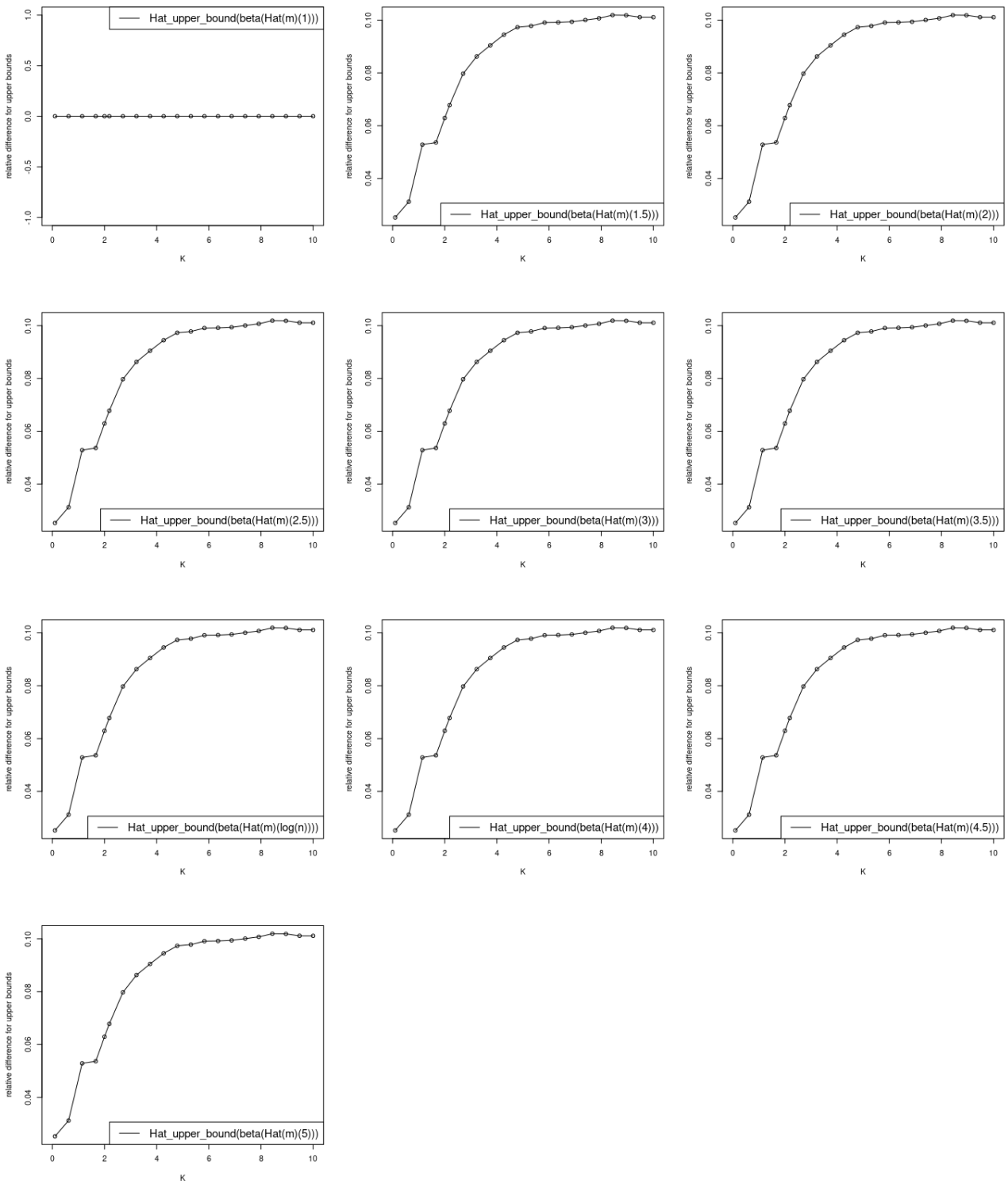


Figure 4.22: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from only one dataset. These plots are obtained with the toy data set described in Subsection 4.3.1 for $\beta^*_{10} = 2$ and distant coefficients.

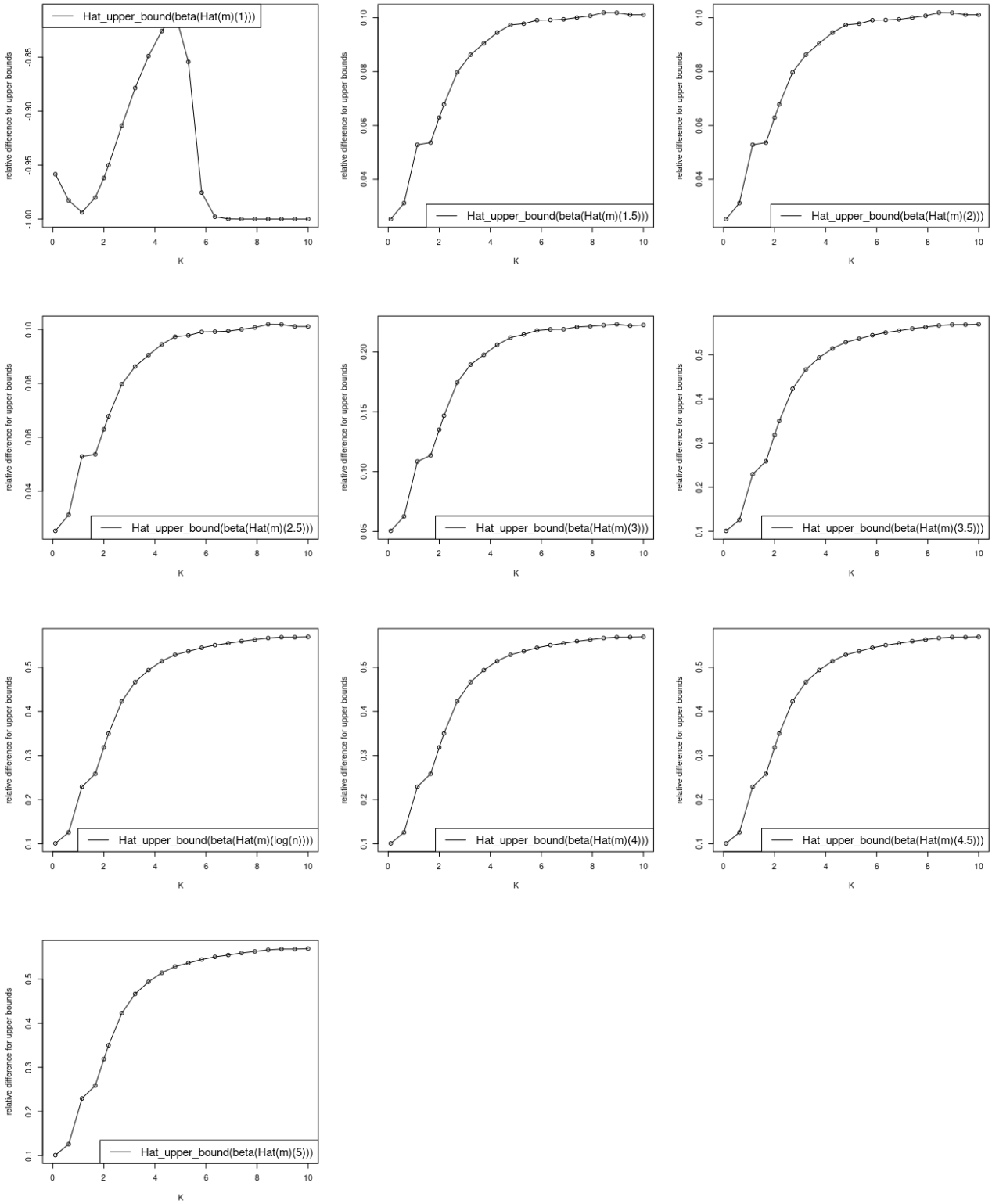


Figure 4.23: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from only one dataset. These plots are obtained with the toy data set described in Subsection 4.3.1 for $\beta^*_{10} = 2$ and close coefficients.

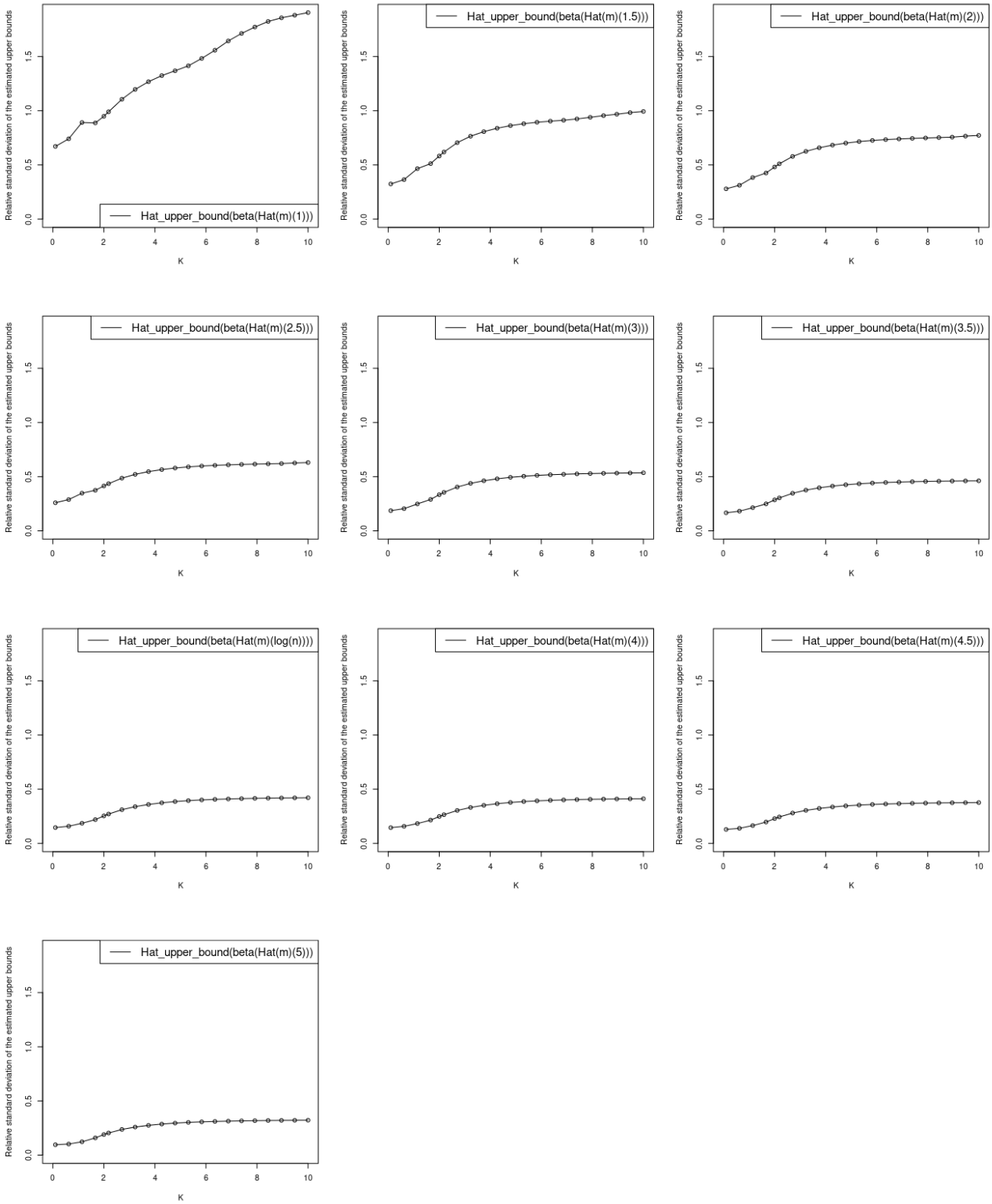


Figure 4.24: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. These plots are obtained with the toy data set described in Subsection 4.3.1 for $\beta^*_{10} = \frac{2}{10}$.

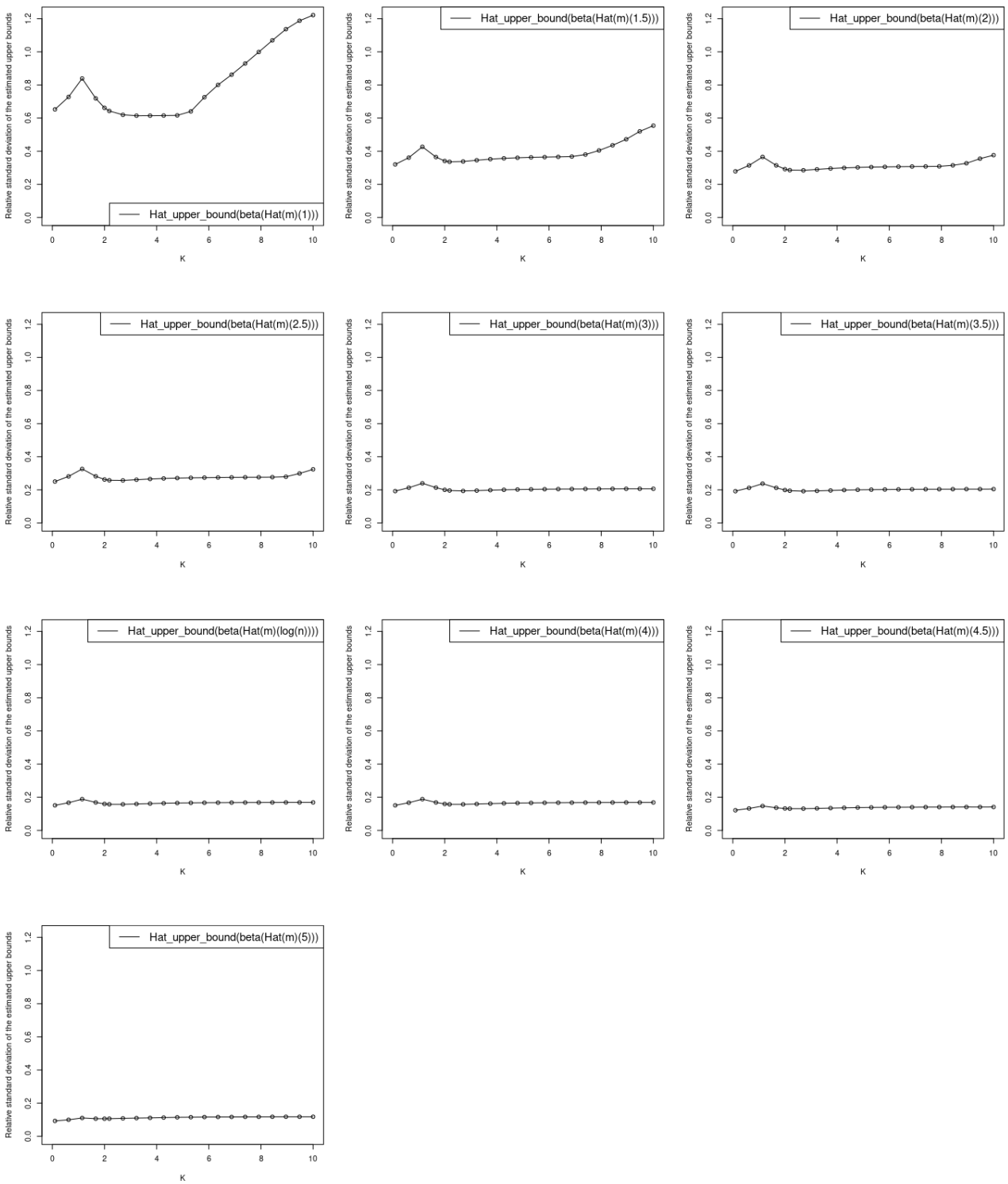


Figure 4.25: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. These plots are obtained with the toy data set described in Subsection 4.3.1 for $\beta^*_{10} = 2$ and distant coefficients.

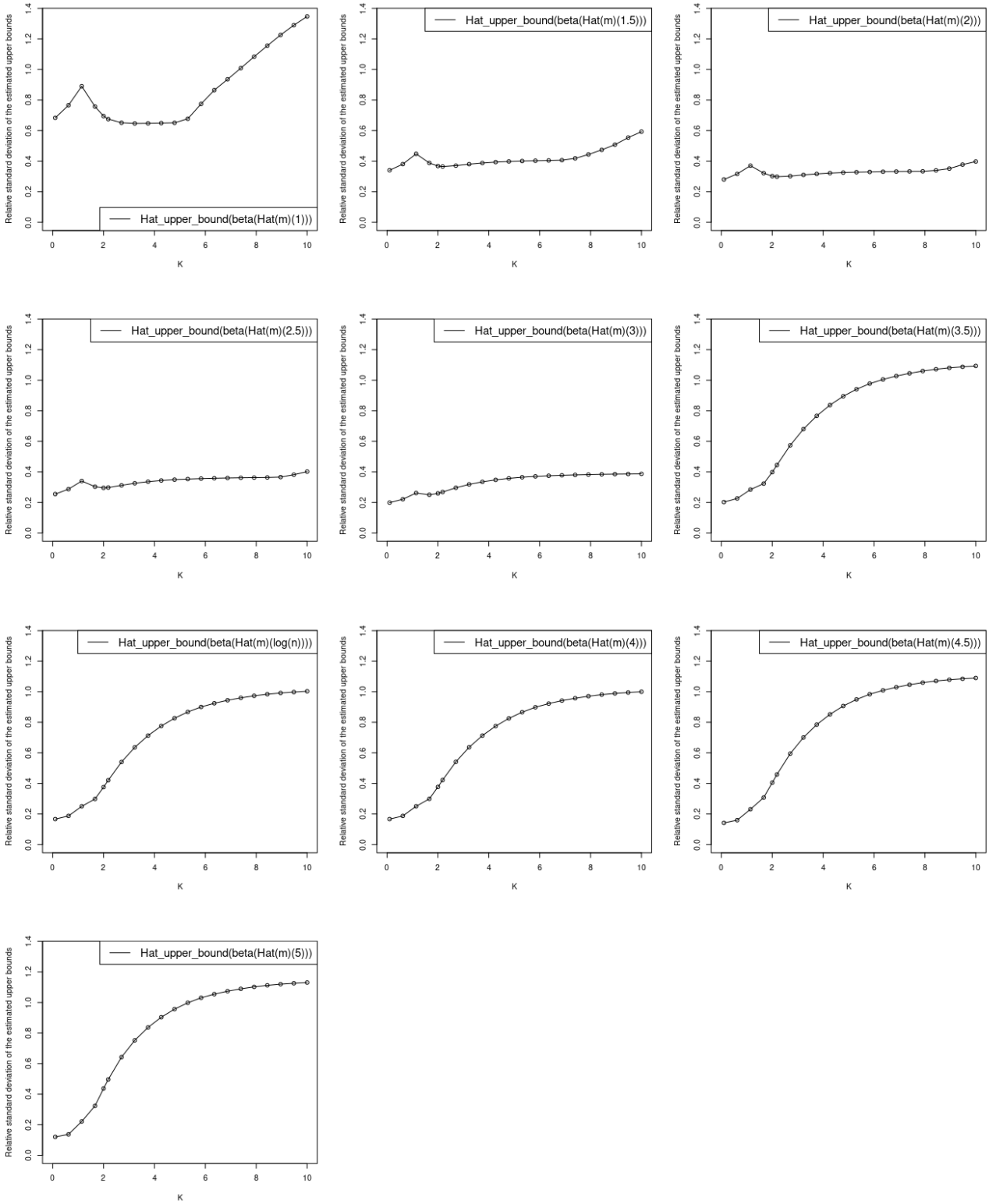
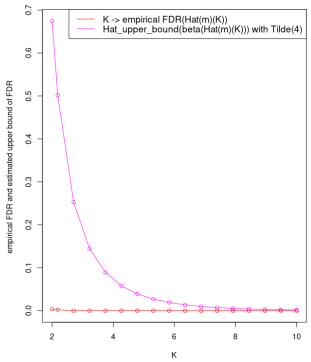
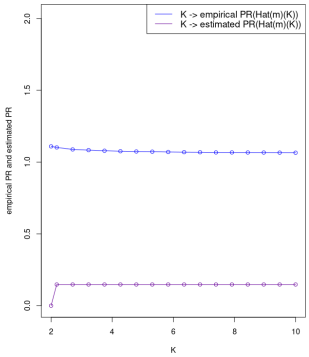
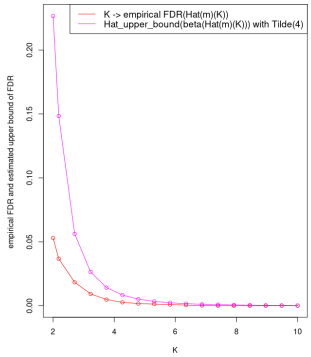


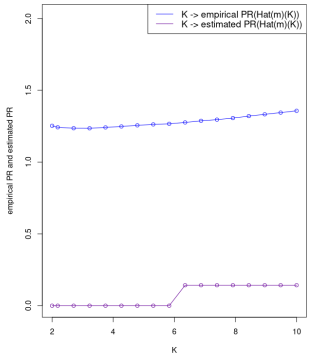
Figure 4.26: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. These plots are obtained with the toy data set described in Subsection 4.3.1 for $\beta^*_{10} = 2$ and close coefficients.



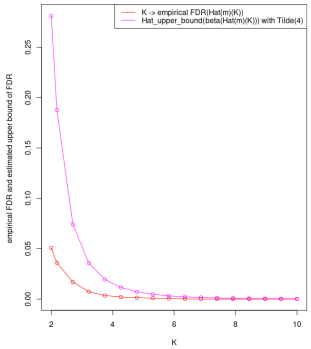


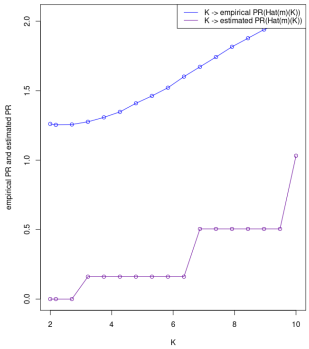
$\beta^*_{10} = \frac{2}{10}$





$\beta^*_{10} = 2$ and distant coefficients





$\beta^*_{10} = 2$ and close coefficients

Figure 4.27: Curves of the empirical estimation functions of $\mathrm{FDR}\big(\hat{m}(K)\big)$ and $\mathrm{PR}\big(\hat{m}(K)\big)$ for all $K > 0$ by using 1000 datasets and curves of the estimated risk (4.22) and the $\tilde{B}(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ function obtained in Corollary 4.3 by replacing β^* by $\hat{\beta}_{\hat{m}(4)}$. These two last plots are obtained from only one dataset. These plots are obtained with the toy data set described in Subsection 4.3.1 for $\beta^*_{10} = \frac{2}{10}$, for $\beta^*_{10} = 2$ and distant coefficients, $\beta^*_{10} = 2$ and close coefficients.

Chapter 5

A 2D-slope heuristics to calibrate hyperparameters in data-driven penalty functions

Abstract

The slope heuristic principle aims at estimating hyperparameters. Two algorithms are extensively used in this setting: the dimension jump and the slope estimation. We propose to study the last one associated with the minimal penalty function defined in 5.7. Two hyperparameters have then to be calibrated. In this Chapter, considering the random model collection, we propose new algorithms. The first is a generalization of the R function *DDSE* of the R package *Capushe*. It is based on a constrained minimization problem to obtain positive estimations of both hyperparameters. The second one consists of adding the model collection's randomness in the first proposed algorithm. It is based on a resampling procedure.

Keywords: Model selection, slope heuristics principle, slope estimation algorithm.

Contents

5.1 Introduction.

5.1.1 Model and problematic.

As in Chapter 3, we consider the following high-dimensional Gaussian linear regression model:

$$Y = X\beta^* + \varepsilon. \tag{5.1}$$

The response vector $Y = (y_i)_{\{1\leq i\leq n\}} \in \mathbb{R}^n$ is regressed on p given vectors denoted $X_1 = (x_{1i})_{\{1\leq i\leq n\}}, \cdots, X_p = (x_{pi})_{\{1\leq i\leq n\}}$. We denote by $X = (X_1, \cdots, X_p)$ the design matrix of size $n \times p$. We suppose that variables are partially correlated with each other. We assume being in the homoscedastic framework: the noise $\varepsilon = (\varepsilon_i)_{\{1\leq i\leq n\}}$ is Gaussian: $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ with an unknown variance $\sigma^2 > 0$.
We assume that p, the dimension of the unknown parameter, is large and is of the order of magnitude or eventually larger than n. In this context of high-dimensional framework, additional regularity assumptions are required. In this work, we assume that β^* is sparse, meaning that only a few coefficients are non-zero comparing to p. This means that a few variables among $(X_1, \cdots, X_p)$ are involved to explain Y. Moreover, to avoid the ultra high-dimensional setting [Verzelen et al., 2012], p can be large but not excessively: if k denotes the number of non-zero coefficients in β^*, then $2k\log(\frac{p}{k}) < n$ is required (see [Giraud et al., 2012]), which is assumed in this article.
The goal is to estimate β^* to ensure accurate prediction performance.

5.1.2 Notations

The notation $\mathcal{D} = (Y, X)$ denotes the available dataset. We define the model m^* as the support of β^*: $m^* = \text{Span}(X_j, j\ s.t.\ \beta_j^* \neq 0)$. A model m is a linear subspace of $\mathbb{R}^p$ corresponding to a sequence of variables X_j for some $j \in \{1, \cdots, p\}$. Hence, $m = \text{Span}\Big((X_j)_{j\in J},\ J \in \mathcal{P}(\{1, \cdots, p\})\Big)$ where $\mathcal{P}(\{1, \cdots, p\})$ designs all the subsets of $\{1, \cdots, p\}$. We denote the dimension of m by D_m and the orthogonal projection onto m by Π_m.
We denote by $||.||_2$ and $||.||_{2,n}$ for respectively the usual Euclidean norm and the usual standard Euclidean norm on $\mathbb{R}^n$. The norms $|.|_1$ is the usual norm 1 on $\mathbb{R}^p$.
The predictive risk of any estimator β is defined by:

$$\mathbb{E}_Y[||X\beta - X\beta^*||_{2,n}^2], \tag{5.2}$$

where $\mathbb{E}_Y$ designs the expectation under the distribution of Y satisfying (5.1). We define the least squares criterion:

$$\forall \beta \in \mathbb{R}^p, \quad \gamma_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2$$

For all positive integers (q, r) and all $W \in \mathcal{M}_{q,r}(\mathbb{R})$, the matrix norm of W is $|||W||| = \sup_{x\in\mathbb{R}^r, x\neq 0}\Big\{\frac{||Wx||_{2,n}}{||x||_{2,n}}\Big\}$. Finally, $\langle .,. \rangle$ designs the canonical scalar product in $\mathbb{R}^n$: $\forall (a, b) \in \mathbb{R}^n$, $\langle a, b\rangle = \sum_{i=1}^{n} a_i b_i$.

5.1.3 Model selection by least-squares penalization.

In the context of high-dimension, we process the variable selection procedure via model selection procedure by penalized criteria.
Let $\mathcal{M}(\mathcal{D})$ be a random model collection built from the available dataset $\mathcal{D}$. This collection is a set of some variables subsets: each $m \in \mathcal{M}(\mathcal{D})$ is a subspace of $\mathbb{R}^p$ spanned by the variables of the associated subset. The least-squares estimators onto model m of $\mathcal{M}(\mathcal{D})$ are defined by:

$$\hat{\beta}_m = \underset{\{\beta, X\beta\in m\}}{\arg\min}\, \gamma_n(\beta), \tag{5.3}$$

given a collection of estimators. The goal of the model selection procedure is to select the best model m among $\mathcal{M}(\mathcal{D})$ leading to an optimal predictive risk control.

- Model selection:
The model selection theory has been developed in [Birgé and Massart, 2001] for a non-asymptotic predictive point of view and when the model collection $\mathcal{M}$ is supposed to be deterministic. The final selected estimator is ideally the one minimizing the predictive risk among the available collection (5.2). Its minimization is impossible because of its dependence on the unknown parameter β^*. In practice, this quantity is replaced by the least-squares function and the criterion to minimize is:

$$\text{crit}(m) = \gamma_n(\hat{\beta}_m) + \text{pen}(m). \tag{5.4}$$

The function pen is a positive and increasing function depending on D_m. The final selected model $\hat{m}$ is defined by:

$$\hat{m} = \underset{m\in\mathcal{M}}{\arg\min}\Big\{\text{crit}(m)\Big\};$$

The main challenge is to get the best trade-off between the sparsity of $\hat{\beta}_{\hat{m}}$ (small number of selected variables) and the goodness of fit (small least-squared values), this through a good choice of penalty function pen.

In this direction, L.Birgé and P.Massart propose a definition of a minimal penalty function $\text{pen}_{\min}$ such that for all penalty functions pen satisfying pen $\geq \text{pen}_{\min}$, the associated selected model verifies the following inequality called an oracle inequality on the predictive risk:

$$\mathbb{E}_Y[||X\hat{\beta}_{\hat{m}} - X\beta^*||_{2,n}^2] \leq C_n \inf_{m\in\mathcal{M}} \mathbb{E}_Y[||X\hat{\beta}_m - X\beta^*||_{2,n}^2] + R_n,$$

with $C_n \approx 1$ at least for large n and R_n is small comparable to $\inf_{m\in\mathcal{M}} \mathbb{E}_Y[||X\hat{\beta}_m - X\beta^*||_{2,n}^2]$.
By assuming that the variance σ^2 is known and the model collection $\mathcal{M}$ is the full exhaustive models, they obtain the following explicit minimal penalty function (Theorem 1. of [Birgé and Massart, 2007]): $\forall m \in \mathcal{M}$,

$$\text{pen}_{\min}(m) = \sigma^2 \frac{D_m}{n}\Big(\kappa + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m\Big). \tag{5.5}$$

where $\{L_m\}_{m\in\mathcal{M}}$ a family of weights, i.e. a family of nonnegative real numbers, satisfying:

$$\Sigma = \sum_{\{m\in\mathcal{M}, D_m>0\}} e^{-L_m D_m} < +\infty. \tag{5.6}$$

Lastly, L.Birgé and P.Massart propose in [Birgé and Massart, 2007] to choice pen $=$ pen_{opt} in (5.4) where:

$$\forall m \in \mathcal{M}, \quad \text{pen}_{\text{opt}}(m) = K\sigma^2 \frac{D_m}{n}\Big(\kappa + 2(2-\theta)\sqrt{L_m} + 2\theta^{-1}L_m\Big), \quad \text{for } \kappa > 2-\theta, \quad \theta \in]0,1[\quad \text{and } K > 1.$$

5.1.4 The data-driven penalties

In our framework (5.1) with a high-dimensional context, an unknown variance and a large model collection, two optimal penalty functions exist for non-asymptotic guarantees on the predictive risk (see Chapter 3 for more details): the LinSelect penalty function [Baraud et al., 2009] and the data-driven penalties [Birgé and Massart, 2007].
According to Chapter 3, the data-driven penalties give worst prediction performances than the LinSelect penalty. The main reasons are that they are not adapted to the Gaussian linear regression model and a random model collection. In this Chapter, we propose new data-driven penalties adapted to this context by generalizing the existing data-driven penalties principle.

The two main ideas of the data-driven penalties are to consider the unknown constants appearing in the penalty as hyperparameters to calibrate them on the available dataset directly and to include the unknown variance σ^2 in these hyperparameters.
After some computations to choice the family of weights $\{L_m\}_{m\in\mathcal{M}(\mathcal{D})}$ adapted in our framework (5.1) (see Subsection 2.2.2 of Chapter 2 for the details of computations), the minimal penalty function is rewritten by:

$$\forall m \in \mathcal{M}(\mathcal{D}), \quad \text{pen}_{\min}(m) = \left(C_1(\sigma^2)\frac{D_m}{n} + C_2(\sigma^2)\frac{\log\left(\binom{p}{D_m}\right)}{n}\right) \tag{5.7}$$

where $C_1(\sigma^2)$ and $C_2(\sigma^2)$ are the two unique unknown constants depending on σ^2.

5.1.5 Objectives

Firstly, we propose a new algorithm based on the slope heuristics method to calibrate the two hyperparameters $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ in the data-dependent penalties. Moreover, the data-driven penalties are data-dependent but theoretical properties are known only for a deterministic model collection (Theorem 1. of [Birgé and Massart, 2007]). So, secondly, we propose to add the randomness of the model collection in the data-driven penalties calibration.
We compare the predictive risk of the selected model from our method with that derived from existing slope heuristics methods and the LinSelect penalty (see Chapter 3 for more details).

5.1.6 Plan of the chapter.

The rest of the chapter is organized as follows. Section 5.2 presents the slope heuristics method in a theoretical point of view (Subsection 5.2.1) and in a practical point of view (Subsection 5.2.2). Section 5.3 is devoted to the presentation of the two existing slope heuristics algorithms (Subsection 5.3.1) and the description of the R function *DDSE* (Subsection 5.3.2). Sections 5.4 and 5.5 contain the contributions of this chapter: after presenting the state of the art of the $2D$-slope estimation principle applications (Subsection 5.4.1), the first contribution is the generalization of slope estimation algorithm to calibrate two hyperparameters (Subsection 5.4.2), the second one consists in adding the collection randomness (Section 5.5). Finally, Section 5.6 stands for a simulation study with firstly the presentation of the simulation framework (Subsection 5.6.1), secondly an analysis of the output characteristics of the proposed algorithms (Subsection 5.6.2) and thirdly, results of the predictive risk evaluations on our proposed methods in comparison with the state of the art (Subsection 5.6.3). A conclusion and some perspectives can be found respectively in Sections 5.7 and 5.8.

5.2 The slope heuristics method for data-dependent penalties

In this section, we present the slope heuristics method in two steps. The first one is based on theoretical consideration (see Subsection 5.2.1) and the second one is based on practical consideration (see Subsection 5.2.2).

5.2.1 In a theoretical point of view

In [Birgé and Massart, 2007], the authors propose a method to calibrate the hyperparameters. This method, called the slope heuristics principle, is based on theoretical aspects and heuristic ideas. This leads to the definition of efficient and tractable pen_{opt} functions that are data-driven. Their framework is the Gaussian linear regression (5.1) with known variance, a fixed design, and a full exhaustive model collection but not in a high-dimensional context.

The rest of this subsection describes the first step of the slope heuristics principle. It consists in studying the theoretical $\text{pen}_{\text{ideal}}$, pen_{opt} and pen_{min} functions defined in Subsection 2.2.2 of Chapter 2 to lead to a definition of a final tractable pen_{opt} available from the dataset.

- The ideal penalty function:
Minimizing the predictive risk (5.2) is equivalent to minimize the expected risk (more details are given in Subsection 2.2.2 of Chapter 2). It is defined by :

$$\forall \beta \in \mathbb{R}^p, \quad \ell(\beta^*, \beta) = \mathbb{E}_Y[||Y - X\beta||_{2,n}^2] - \mathbb{E}_Y[||Y - X\beta^*||_{2,n}^2]. \tag{5.8}$$

From this last definition (5.8):

$$\begin{aligned}
\forall m \in \mathcal{M}, \qquad \ell(\beta^*, \hat{\beta}_m) - \gamma_n(\hat{\beta}_m) &= \mathbb{E}_Y[||Y - X\hat{\beta}_m||_{2,n}^2] - \mathbb{E}_Y[||Y - X\beta^*||_{2,n}^2] - \gamma_n(\hat{\beta}_m) \\
&= \mathbb{E}_Y[||Y - X\hat{\beta}_m||_{2,n}^2] - \mathbb{E}_Y[||Y - X\beta^*||_{2,n}^2] - \gamma_n(\hat{\beta}_m) \\
&\quad - \mathbb{E}_Y[||Y - X\beta_m||_{2,n}^2] + \mathbb{E}_Y[||Y - X\beta_m||_{2,n}^2] \\
&\quad - \gamma_n(\beta_m) + \gamma_n(\beta_m) \\
&\quad - \gamma_n(\beta^*) + \gamma_n(\beta^*) \\
&\underset{(**)}{=} \mathbb{E}_Y[||Y - X\hat{\beta}_m||_{2,n}^2] - \mathbb{E}_Y[||Y - X\beta_m||_{2,n}^2] \\
&\quad + \gamma_n(\beta_m) - \gamma_n(\hat{\beta}_m) \\
&\quad + \mathbb{E}_Y[||Y - X\beta_m||_{2,n}^2] - \mathbb{E}_Y[||Y - X\beta^*||_{2,n}^2] \\
&\quad + \gamma_n(\beta^*) - \gamma_n(\beta_m) - \gamma_n(\beta^*) \\
&\underset{(***)}{=} v_m + \hat{v}_m + \Delta_n(\beta_m) - \gamma_n(\beta^*)
\end{aligned}$$

with:

1. $v_m = \mathbb{E}_Y[||Y - X\hat{\beta}_m||_{2,n}^2] - \mathbb{E}_Y[||Y - X\beta_m||_{2,n}^2]$ is a variance term
2. $\hat{v}_m = \gamma_n(\beta_m) - \gamma_n(\hat{\beta}_m)$ is an empirical variance term
3. $\Delta_n(\beta_m) = \mathbb{E}_Y[||Y - X\beta_m||_{2,n}^2] - \mathbb{E}_Y[||Y - X\beta^*||_{2,n}^2] + \gamma_n(\beta^*) - \gamma_n(\beta_m)$ is a residual term
4. $-\gamma_n(\beta^*)$ is a term independent of $m \in \mathcal{M}$.

(**) is just a reordering of terms and *(***)* is true since $X\hat{\beta}_m = \Pi_m(Y)$, which leads to $X\hat{\beta}_{\hat{m}} - Y \in m^\perp$.
Since the term $-\gamma_n(\beta^*)$ does not depend on $m \in \mathcal{M}$, minimizing $\text{pen}^*_{\text{ideal}}$ with respect to $m \in \mathcal{M}$ is equivalent to minimize the function

$$\text{pen}_{\text{ideal}}(m) = v_m + \hat{v}_m + \Delta_n(\beta_m), \quad \forall m \in \mathcal{M}.$$

- The minimal penalty function:
In $\text{pen}_{\text{ideal}}(m) = v_m + \hat{v}_m + \Delta_n(\beta_m)$, the only term available is the empirical variance $\hat{v}_m$. Let

us study the selected model $\hat{m}$ when $\text{pen}(m) = \kappa \hat{v}_m$ for a positive κ. The criterion (5.4) to minimize becomes:

$$\text{crit}(m) = \gamma_n(\hat{\beta}_m) + \kappa \hat{v}_m = (1-\kappa)\gamma_n(\hat{\beta}_m) + \kappa \gamma_n(\beta_m).$$

If $\kappa = 1$, $\text{crit}(m) = \gamma_n(\beta_m)$, decreasing when D_m increases. That means that the bias is only the remained term and the variance is not taken into account anymore; and as when n is large, $\text{crit}(m) \approx \mathbb{E}_Y[||Y - X\beta_m||_{2,n}^2]$, then $\hat{m}$ is the one minimizing the bias and so, $D_{\hat{m}}$ tends to be huge. In the same way, as the bias is almost constant for the large models, both terms in the crit function decrease when the dimension increases if $\kappa < 1$. So, the dimension of $\hat{m}$ is large. On the contrary, for large models, the crit function increases with respect to model dimension if $\kappa > 1$. So, models with large dimensions are not selected and $\hat{m}$ has a reasonable dimension. Hence, there is a transition phase in $\kappa = 1$ meaning that for all $\kappa > 1$, $D_{\hat{m}}$ is reasonable whereas for all $\kappa < 1$, $D_{\hat{m}}$ explodes. Hence, $[m \mapsto \hat{v}_m]$ is the minimal penalty function in the sense that all pen functions such that $\text{pen}(m) > \text{pen}_{\text{min}}(m)$ lead to a reasonable model dimension whereas all pen functions such that $\text{pen}(m) < \text{pen}_{\text{min}}(m)$ lead to models with too large dimensions and exploding values of predictive risk. So, the minimal penalty function is close to $\hat{v}_m$, $\forall m \in \mathcal{M}$.

- The optimal penalty function:
According to the Theorem 1. of [Birgé and Massart, 2007], to select a model with a reasonable dimension, the pen_{opt} function has to satisfy $\text{pen}_{\text{opt}}(m) \geq \text{pen}_{\text{min}}(m)$ for all $m \in \mathcal{M}$, with $\text{pen}_{\text{min}}(m) \approx= \hat{v}_m$. Moreover, as explained in Subsection 2.2.2 of Chapter 2, to derive an oracle inequality, $\text{pen}_{\text{opt}}(m) \approx \text{pen}_{\text{ideal}}(m)$, and $\text{pen}_{\text{ideal}}(m) \approx v_m + \hat{v}_m + \Delta_n(\beta_m)$, $\forall m \in \mathcal{M}$. So, the goal is to find a pen_{opt} function, tractable, close to $\text{pen}_{\text{ideal}}$ and larger than the $\hat{v}_m$'s values.
Let us study the $\text{pen}_{\text{ideal}}$ function. The term $\hat{v}_m$ is a consistent estimator of v_m. Hence, $\hat{v}_m \approx v_m$ for m large. Concerning the $\Delta_n(\beta_m)$ term, when D_m is large, it is concentrated around its expectation, which is zero (by concentration inequality). So, $\Delta_n(\beta_m)$ is small for large D_m. Hence, the approximation $\text{pen}_{\text{ideal}}(m) \approx 2\hat{v}_m$ is valid when D_m is large. In this way, a tractable pen_{opt} function, at least for D_m large, is $\text{pen}_{\text{opt}}(m) = 2\hat{v}_m$, $\forall m \in \mathcal{M}$, which is an approximation of $2\ \text{pen}_{\text{min}}(m)$.

By combining the minimal penalty function (5.7) and $\text{pen}_{\text{opt}} \approx 2\ \text{pen}_{\text{min}}$, the model selection procedure in our framework is applied by minimizing (5.4) with the optimal penalty function defined by:

$$\text{pen}_{\text{opt}}(m) = 2\left(C_1(\sigma^2)\frac{D_m}{n} + C_2(\sigma^2)\frac{\log(\binom{p}{D_m})}{n}\right), \tag{5.9}$$

where $C_1(\sigma^2)$ and $C_2(\sigma^2)$ are the two unique unknown constants depending on the experimental design only through σ^2.

5.2.2 In a practical point of view

In our context, the tractable optimal penalty function is (5.9) depending on constants. From a practical point of view, the slope heuristic considers that these unknown constants are hyperparameters to calibrate from the dataset. We present in the sequel the second part of the

slope heuristics principle. It has been introduced in [Birgé and Massart, 2007] to overcome the problem of σ^2 unknown which is included in the hyperparameters to be calibrated. It is based on a mixture of theoretical and heuristic ideas.

We recall that for all $m \in \mathcal{M}$:

- $\text{pen}_{\text{opt}}(m) = 2\hat{v}_m$
- $\text{pen}_{\text{min}}(m) = \hat{v}_m \approx \left(C_1(\sigma^2)\frac{D_m}{n} + C_2(\sigma^2)\frac{\log(\binom{p}{D_m})}{n} \right)$

We define the function $\text{pen}_{\text{shape}}$ such that pen_{min} is a linear combination of $\text{pen}_{\text{shape}}$:

$$\forall m \in \mathcal{M}, \quad \text{pen}_{\text{min}}(m) = \langle C^T, \text{pen}^T_{\text{shape}}(m) \rangle,$$

where $C \in \mathbb{R}^q$ is the vector of hyperparameters, for a certain q, and $\text{pen}_{\text{shape}}$ is a function with q inputs.
With the definition of pen_{min} in (5.7), $C = \Big(C_1(\sigma^2), C_2(\sigma^2) \Big)$, the corresponding $\text{pen}_{\text{shape}}$ is:

$$\forall m \in \mathcal{M}, \quad \text{pen}_{\text{shape}}(m) = \Big(\frac{D_m}{n}, \frac{\log(\binom{p}{D_m})}{n} \Big)$$

Moreover:

$$\begin{aligned} \forall m \in \mathcal{M}, \quad \hat{v}_m &= \gamma_n(\beta_m) - \gamma_n(\hat{\beta}_m) \\ &= \gamma_n(\beta_m) - \gamma_n(\beta^*) + \gamma_n(\beta^*) - \gamma_n(\hat{\beta}_m). \end{aligned}$$

The first term $\gamma_n(\beta_m) - \gamma_n(\beta^*)$ is the empirical bias term and is stable when D_m is large enough. In the second term, $\gamma_n(\beta^*)$ does not depend on m. Hence, for D_m large enough:

$$\hat{v}_m \underset{D_m \text{ large}}{\approx} -\gamma_n(\hat{\beta}_m) + \eta,$$

with η a real constant independent of $m \in \mathcal{M}$. Hence, the behavior of $[m \mapsto \hat{v}_m]$ is known for D_m large enough. Indeed, the $-\gamma_n(\hat{\beta}_m)$ function is an affine one with respect to the $\text{pen}_{\text{min}}(m) = \hat{v}_m$ values, and the slope is equal to 1. The pen_{min} is thus the function that compensates the least-squared values and the pen_{opt} function, larger than the pen_{min} one, avoids the selection of too large models. Hence, with the $\text{pen}_{\text{shape}}$ definition and to summarize the second step of the slope heuristics principle, hyperparameters $\Big(C_1(\sigma^2), C_2(\sigma^2) \Big)$ of the pen_{min} function are tuned as the slope coefficients estimation of the affine behavior between the values of $-\gamma_n(\hat{\beta}_m)$ and the $\text{pen}_{\text{shape}}$ ones for D_m large enough. The pen_{min} function is a data-driven penalty function since it is calculated directly from the data. So, the slope heuristics principle leads to a flexible penalty function and is adapted to the unknown variance framework since σ^2 is tuned during the hyperparameters calibration. For more details of the slope heuristics method, we refer the readers to [Arlot, 2019].

5.3 Slope estimation: calibration of only one hyperparameter

The slope heuristics principle is an automatic method to calibrate the unknown hyperparameters in optimal penalty functions for model selection. The advantage of this method is to be data-driven.

5.3.1 Two algorithms for the slope heuristics principle

Two algorithms are available when the optimal penalty function is known up to a unique hyperparameter: *the dimension jump algorithm* and *the data-driven slope estimation algorithm.* Both algorithms derive from the slope heuristic principle and offer two different visualizations of applying the model selection procedure on data sets.

The dimension jump algorithm.
The main idea for the dimension jump algorithm is that too small penalty functions lead to selecting a large model dimension with a high predictive risk. In contrast, for the other penalty functions, the dimension of the selected model becomes reasonable. The dimension jump algorithm can be summarized as follows:

Step 1: Compute

$$\forall \kappa > 0, \quad m(\kappa) \in \underset{m \in \mathcal{M}(\mathcal{D})}{\arg\min} \left\{ \gamma_n\left(\hat{\beta}_m\right) + \kappa \ \mathrm{pen}_{\mathrm{shape}}(m) \right\}$$

Step 2: Find $\hat{\kappa}$ such that dimension of $m(\kappa)$ is large when $\kappa < \hat{\kappa}$ and is reasonable otherwise.

Step 3: Select $\hat{m} = m(2\hat{\kappa})$.

The slope estimation algorithm.
The slope heuristic algorithm, also called the *Slope Heuristic Robust Regression* in the literature, is mainly based on the following key idea: the unknown hyperparameter is directly calibrated by the slope of the expected theoretical affine behavior between the $-\gamma_n\left(\hat{\beta}_m\right)$ values and the penalty shape values for large models. From the cloud formed by the values of the couples $\left(\mathrm{pen}_{\mathrm{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ of the model collection, the best linear trend in one dimension is researched to get a slope. Then, by removing one by one the couples with the smallest penalty shape value:

Step 1: Compute κ which is the estimation of the slope of the best affine trend in one dimension corresponding to the remaining couples of points

Step 2: Compute $m(\kappa) = \underset{m \in \mathcal{M}(\mathcal{D})}{\arg\min} \left\{ \gamma_n \left(\hat{\beta}_m \right) + \kappa \mathrm{pen}_{\mathrm{shape}}(m) \right\}$

Step 3: Plot the function $m(\kappa)$ with respect to the number of couples considered for $m(\kappa)$, and find the last large plateau

Step 4: Select $\hat{m} = m(2\hat{\kappa})$

Step 3 is the stabilization of the selected model on the affine behavior: if the remaining models are only those corresponding to the affine behavior, then, by removing them one by one, the slope is almost constant, as well as the selected model.

For the sequel, we decide to focus on the slope estimation algorithm. The main reason is that the dimension jump algorithm is based on a sudden fall of the dimension of $m(\kappa)$ when κ increases. However, several large jumps are most of the time observed. To circumvent this problem, solutions choose the largest one or the first brutal change of dimensions, which requires the definition of severe change of dimensions. Unfortunately, the selected model usually differs according to these two solutions. On the contrary, the slope estimation algorithm is based on an affine relation which is unique for large models. Moreover, apart from in [Devijver and Gallopin, 2018] where the dimension jump algorithm is seen as similar to the slope estimation one, the other slope heuristics applications showed that the dimension jump algorithm is less reliable than the slope heuristics one.

5.3.2 Description of the slope estimation algorithm (R function *DDSE*)

The slope heuristic algorithm is implemented in the The R function *DDSE* (for *Data-Driven Slope Estimation*) of the R package *Capushe* (for *Calibration of penalized criteria for model selection*). To directly estimate the slope from the expected affine behavior between the values $-\gamma_n \left(\hat{\beta}_m \right)$ and the penalty for large models, two challenges have to be considered:

1. How to select the couples $\left(\mathrm{pen}_{\mathrm{shape}}(m), -\gamma_n \left(\hat{\beta}_m \right) \right)$ with large $\mathrm{pen}_{\mathrm{shape}}(m)$ values and satisfying the expected theoretical affine relation?
2. How to obtain a suitable estimation of the slope given a reasonable hyperparameter calibration?

Step 1:
The first step consists in some modifications on the model collection leading to a collection

with almost at most one model per dimension.

Step 2:
The second step of the R function *DDSE* of the R package *Capushe* consists in searching for the best linear trend in one dimension in the clouds formed by the values of couples $\left(\text{pen}_{\text{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ of the model collection. This research gives the line parameters, and especially the slope denoted $\kappa(\sigma^2)$. This slope $\kappa(\sigma^2)$ gives an available penalty function and a selected model can be computed by:

$$\hat{m}(\kappa(\sigma^2)) = \underset{m\in\mathcal{M}(\mathcal{D})}{\arg\min}\left\{\gamma_n\left(\hat{\beta}_m\right) + 2\kappa(\sigma^2)\text{pen}_{\text{shape}}(m)\right\}. \tag{5.10}$$

Then, the algorithm performs the same procedure removing one by one the couple $\left(\text{pen}_{\text{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ with the smallest $\text{pen}_{\text{shape}}(m)$ value, this until there only remains one pair of values $\left(\text{pen}_{\text{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ At the end of this step, the function of the successive selected models with respect to the number of couples $\left(\text{pen}_{\text{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ used for the linear regression coefficient estimation is available. The main key idea of the R function *DDSE* is the stabilization of the slope when we consider at most the couples $\left(\text{pen}_{\text{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ corresponding to the affine behavior if the remaining couples are only those corresponding to the affine behavior, then, by removing them one by one, the slope is almost constant. So, in this area, the penalty values remain constant, as well as the selected models. Thus, the last step of the algorithm is based on the stabilization of the selected models. More precisely, the function of the successive selected models with respect to the number of couples $\left(\text{pen}_{\text{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ used for the linear regression coefficient estimation is piecewise constant.

Step 3:
The selected plateau is the one of a length greater than a certain threshold denoted "*pct*". This threshold is chosen beforehand by the user. This plateau corresponds to an estimated $\hat{\kappa}(\sigma^2)$ and the final selected model $\hat{m}(\hat{\kappa}(\sigma^2))$ minimizes (5.10) on the entire collection with $\kappa(\sigma^2) = \hat{\kappa}(\sigma^2)$. One of the main advantages of the R function *DDSE* is that the algorithm is not too sensitive to the parameter "*pct*" thanks to the use of robust linear regression. Indeed, the *rlm* R function (for *robust fitting of linear models*), is used, allowing it to be robust to outliers. This function is based on an iterated algorithm solving re-weighted least squares minimization: it used an extension of the M-estimator instead of the ordinary least-squares one. More precisely, we propose to use the bi-squared function of Tukey for the M-estimator. At each iteration, weights are added to each observation. There are chosen to be inversely proportional to the variance within the points cloud so that the extreme points are weighted by a too small value so that they have very little weight when searching for the line passing

through the point cloud. Consequently, in addition to remove outlier models, especially for the largest models of the collection, when estimation can be noisy due to the complexity of the $\hat{\beta}_m$ computation, the robustness process removes the couples $\left(\mathrm{pen}_{\mathrm{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ of the affine behavior with a few other couples since they are detected as outliers for the linear relation. The slope estimation is the same as previously as well as the selected model step. Thus, the plateau of the function of the selected models with respect to the number of couples used for the linear regression coefficient estimation, which corresponds to the affine behavior, tends to be large and so easier to detect. So, the selection of the plateau is not too sensitive to the user choice of "*pct*".

Here is the detailed algorithm of the R function *DDSE* of the R package *Capushe*. The presentation of the algorithm includes R language. The notations data$pen and data$LS represent respectively the vector of the penalty shape values and the vector of the least-squares values $\gamma_n\left(\hat{\beta}_m\right)$ along the model collection.

Algorithm 2: R function *DDSE* of the R Package *Capushe*

Data: data; pct
Result: $\hat{m}_{\hat{\kappa}}$; $\hat{\kappa}$

Step 1: On the model collection

```
data = data[-which(is.na(data), arr.ind = TRUE)[, 1], ] ;
# removing the models from the collection where at least one information is a NA
  value ;

data = data[order(data$pen, data$LS), ] ;
# Ranking the model with respect to the dimension ;
# If several models in the collection have the same penalty shape value, keeping only
  the one with the smallest least-squares value ;

plength = length(data$pen);
for i in plength : 2 do
    if data$pen[i] == data$pen[i - 1] then
        data = data[-i, ];
    end
end
```

Step 2: Estimation of all the slopes

```
couplepen = data$pen;
plength = length(couplepen)-1;
couplecontrast = -data$LS;
kappa = numeric(plength);
mhat = numeric(plength);
for p in 1 : plength do
    # robust regression method to estimate the slope;
    # estimation of the slope and the intercept of the best line passing trought the point
      clouds; only the slope is conserved;

    kappa[p] = rlm(couplecontrast ~ couplepen, psi=psi.bisquare)$coefficients[2] ;
    # at each step, the couple with the smallest penalty shape value is removed;

    couplepen = couplepen[-1] ;
    couplecontrast = couplecontrast[-1] ;
    mhat[p] = which.min(data$LS + 2.kappa[p].data$pen)
end
```

Algorithm 2: R function *DDSE* of the R Package *Capushe*

```
Step 3: Locating the affine behavior
Pi = c(1);
N = c();
number = 1;

for p in 2 : plength do
    if mhat[p] != mhat[p - 1] then
        Pi = c(Pi, p)            # increasing sequence of change points;
        N = c(N, number)         # length of plateaus;
        number = 0;
    end
    number = number + 1
end

N = c(N, number);
Pi = c(Pi, p);
Imax = length(N);

while N[Imax] < (pct*sum(N)) do
    Imax = Imax - 1 ;
    # research to the first last plateau with a sufficient large length;
end

p = Pi[Imax] + floor(N[Imax]/2) ;
# the final selected model corresponds to the selected plateau;
m̂_κ̂ = mhat[p];

# robust regression method to estimate the slope;
# estimation of the slope and the intercept of the best line passing through the points
  cloud; only the slope is conserved;

κ̂ = rlm(-data$LS[p:(plength + 1)] ~ data$pen[p:(plength + 1)], psi =
  psi.bisquare)$coefficients[2]
```

5.4 Slope estimation: calibration of two hyperparameters

5.4.1 State of the art: be reduced to the $1D$-slope estimation

The slope heuristics principle of [Birgé and Massart, 2007] has already been studied in practice for different frameworks and with two unknown hyperparameters. We present in this subsection some of these applications in a non-exhaustive way. The first use was proposed in [Lebarbier, 2005] in the context of multiple change-points detection in the mean of the Gaussian process with a full exhaustive variable selection. Under this framework, the author determines an

optimal penalty function following Theorem 2. of [Birgé and Massart, 2001]:

$$\mathrm{pen}_{\mathrm{E.L}}(m) = 2\frac{D_m}{n}\left(C_1(\sigma^2) + C_2(\sigma^2)\log\left(\frac{n}{D_m}\right)\right)$$

where $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ are two unknown hyperparameters depending on σ^2. By assuming in a first step that σ^2 is known, the authors show, from a large experimental study, that the couple $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ can be fixed to $\sigma^2 \times (C_1, C_2)$ where C_1 and C_2 are two hyperparameters now completely independent of the model parameters; and the ratio $\frac{C_1}{C_1}$ can be fixed to 2.5. It is very satisfying since the optimal penalty can be rewritten in a simpler form:

$$\mathrm{pen}_{\mathrm{E.L}}(m) = 2\sigma^2 C_2 \frac{D_m}{n}\left(2.5 + \log\left(\frac{n}{D_m}\right)\right) \tag{5.11}$$

Thus, there is only one hyperparameter left to tune: $\kappa(\sigma^2) = \sigma^2 C_2$. The authors calibrate it using the dimension jump algorithm described in Subsection 5.3.1. This data-driven penalty is included in the comparison study of Chapter 3. These prediction performances are worse than those from the LinSelect penalty. The main reason is probably that the ratio 2.5 was established in a change point detection context and is no longer valid in Gaussian linear regression with a high-dimensional context. Moreover, the ratio 2.5 is probably not adapted to a random and data-dependent model collection.

A second application of the slope heuristics principle can be found in [Meynet and Maugis-Rabusseau, 2012] for model-based clustering in a high-dimensional context by considering a properly generated data-dependent sub-collection of models. In this framework, an optimal penalty tuned according to Theorem 1. of [Birgé and Massart, 2007] is:

$$\mathrm{pen}_{\mathrm{M.M}}(m) = 2\frac{D_m}{n}\left(C_1(\sigma^2) + C_2(\sigma^2)\log\left(\frac{p}{D_m}\right)\right) \tag{5.12}$$

where $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ are two unknown hyperparameters. The authors use twice the slope estimation algorithm described in Subsection 5.3.2 in two successively steps. Firstly, $C_2(\sigma^2)$ is estimated by exploiting the values of $-\gamma_n\left(\hat{\beta}_m\right)$ viewed as an affine function of $\log\left(\frac{p}{D_m}\right)$ ones for large models; secondly, $C_1(\sigma^2)$ is estimated by exploiting the values of $-\gamma_n\left(\hat{\beta}_m\right) - C_2(\sigma^2)\log\left(\frac{p}{D_m}\right)$ viewed as an affine function of $\frac{D_m}{n}$ ones for large models.

In the same context of [Meynet and Maugis-Rabusseau, 2012] (high-dimension, random sub-collection), the authors of [Devijver, 2017] and [Devijver and Gallopin, 2018] apply the slope heuristics principle for respectively Gaussian mixture models and Gaussian graphical models. The optimal penalty is respectively:

$$\mathrm{pen}_{\mathrm{Dev}}(m) = 2\frac{D_m}{n}\left(C_1(\sigma^2) + C_2(\sigma^2)\log\left(\frac{D_m}{n}\right)\right)$$

and

$$\mathrm{pen}_{\mathrm{D.G}}(m) = 2\frac{D_m}{n}\left(C_1(\sigma^2) + C_2(\sigma^2)\log\left(\frac{p(p-1)}{D_m}\right)\right)$$

where $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ are the two unknown hyperparameters. However, for an applied perspective, both articles restrict the optimal penalty function to $\mathrm{pen}_{\mathrm{opt}}(m) = 2C_1(\sigma^2)\frac{D_m}{n}$. The remaining unknown hyperparameter is calibrated by algorithms described in Section 5.3. In [Devijver and Gallopin, 2018], the authors also apply the slope heuristic algorithms for $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ by the two successively steps described in [Meynet and Maugis-Rabusseau, 2012]. The authors highlight the similarity of the performances of these two slope heuristics applications they obtain on their simulated data sets.

Let us cite an application presented in [Arlot et al., 2019] for change-point detection in regression for a non-parametric case with a full exhaustive variable selection. In their paper, the noise level is not constant and its distribution is not Gaussian anymore leading to the use of kernel methods for the least-squares computations. Their optimal penalty function is:

$$\mathrm{pen}_{\mathrm{A.C.H}}(m) = 2\left(C_1(\sigma^2)\frac{D_m}{n} + C_2(\sigma^2)\frac{\log\left(\binom{n-1}{D_m-1}\right)}{n}\right) \tag{5.13}$$

where $\left(C_1(\sigma^2), (\sigma^2)\right)$ are two unknown hyperparameters. Once again, the unknown hyperparameters are estimated by applying the slope heuristics twice successively.

Except for penalty (5.13) depending on binomial coefficients, all the previous minimal penalty functions involve the fractional terms instead of the binomial coefficients, which is simpler numerically.
By using the inequality $\left(\frac{p}{D}\right)^D \leq \binom{p}{D} \leq \left(\frac{ep}{D}\right)^D$, for $D \leq p$, we have: $\forall m \in \mathcal{M}(\mathcal{D})$,

$$\frac{D_m}{n}\left(C_1(\sigma^2) + C_2(\sigma^2)\log\left(\frac{p}{D_m}\right)\right) \leq \mathrm{pen}_{\min}(m) \leq \frac{D_m}{n}\left(C_1(\sigma^2) + C_2(\sigma^2)\left(1 + \log\left(\frac{p}{D_m}\right)\right)\right),$$

where $\mathrm{pen}_{\min}$ is defined in (5.7). When $D_m = o(p)$, then $\frac{D_m}{n} = \underset{D_m \ll p}{o}\left(\frac{D_m}{n}\log\left(\frac{p}{D_m}\right)\right)$. So, $\mathrm{pen}_{\min}(m) \underset{D_m \ll p}{\approx} \frac{D_m}{n}\left(C_1(\sigma^2) + C_2(\sigma^2)\log\left(\frac{p}{D_m}\right)\right)$ and all the previous minimal penalties are equivalent modulo the replacement of the constant p in (5.7) by the number of parameters to estimate in each considered model. When D_m is not negligible with respect to p, all the previous minimal penalties except for (5.13) are smaller than our minimal penalty (5.7). Hence, our minimal penalty, as well as the penalty (5.13), provides a finer control of the predictive risk, selects a smaller model and avoids over-fitting.

To summarize, the slope heuristics principle has often been applied to different models with more or less restrictive assumptions. We refer the readers to [Arlot, 2019] for a detailed survey of the slope heuristics method. However, two unknown hyperparameters have never been estimated simultaneously: either one of them is omitted; or they are estimated successively meaning that the calibration of the second hyperparameter uses the first calibrated hyperparameter.

We propose an algorithm to calibrate both hyperparameters simultaneously. For this purpose, we are inspired by the R function *DDSE* of the R package *Capushe*. We present two generalizations. The first stands for the extension to the estimation of two hyperparameters; the second stands for taking into account the randomness of the collection.

5.4.2 The proposed algorithm to calibrate two hyperparameters

The main objective is to generalize the R function *DDSE* of the R package *Capushe* to calibrate the two hyperparameters $C_1(\sigma^2)$ and $C_2(\sigma^2)$ simultaneously. Here, the $\text{pen}_{\text{shape}}$ values are decomposed into two components: the $\left(\frac{D_m}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ ones and the $\left(\frac{\log\left(\binom{p}{D_m}\right)}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ ones. It is possible to use the R *DDSE* code with the R function *rlm* in 2-dimensions: if data\$pen1 and data\$pen2 design respectively the values of $\left(\frac{D_m}{n}\right)$ and the values of $\left(\frac{\log\left(\binom{p}{D_m}\right)}{n}\right)$, then, the command *rlm(-data\$LS ∼ (data\$pen1 + data\$pen2), psi = psi.bisquare)* gives a couple $\left(\widehat{C_1}(\sigma^2), \widehat{C_2}(\sigma^2)\right)$. However, $\widehat{C_1}(\sigma^2)$ or $\widehat{C_2}(\sigma^2)$ is often negative which is undesirable from the theoretical point of view and $|\widehat{C_1}(\sigma^2)|$ and $|\widehat{C_1}(\sigma^2)|$ are very high, so the $\gamma_n\left(\hat{\beta}_m\right)$ values have no influence; the penalty function values are so often negative which is undesirable from the theoretical point of view; and the predictive risk values are very large. To overcome these difficulties, we propose to rewrite the hyperparameters calibration problem as a constrained minimization problem.

5.4.2.1 Minimization under positive constraints.

The objective is to find the best affine trend in two dimensions for the cloud formed by the couples $\left(\frac{D_m}{n}, \frac{\log\left(\binom{p}{D_m}\right)}{n}, -\gamma_n\left(\hat{\beta}_m\right)\right)_{m\in\widetilde{\mathcal{M}(\mathcal{D})}}$, for all subsets $\widetilde{\mathcal{M}(\mathcal{D})}$ of $\mathcal{M}(\mathcal{D})$, given parameters of the plane: the slope coefficients C_1 and C_2 and the intercept C_3: the goal is to determine $\left(C_1, C_2, C_3\right)$ minimizing

$$\sum_{m\in\widetilde{\mathcal{M}(\mathcal{D})}} \left(-\gamma_n\left(\hat{\beta}_m\right) - \left(C_1\frac{D_m}{n} + C_2\frac{\log\left(\binom{p}{D_m}\right)}{n} + C_3\mathbb{1}_{\#\{m\in\widetilde{\mathcal{M}(\mathcal{D})}\}}\right)\right)^2, \tag{5.14}$$

where $\mathbb{1}_d$ designs the vector of size d composed only of 1. The constraints are $C_1 > 0$ and $C_2 > 0$ and we can also impose $C_3 < 0$ since for negative z-axis $-\gamma_n\left(\hat{\beta}_m\right)$ values and positives slope coefficients, the intercept is necessarily negative.
The following proposition rewrites the problem (5.14) as a minimization problem under constraints. The last one admits a solution and it is unique.

Proposition 5.1. *Let us define the notations:* $y = \left(-\gamma_n\left(\hat{\beta}_m\right)\right)_{m\in\widetilde{\mathcal{M}(\mathcal{D})}}$, $x = \left(\frac{D_m}{n}\right)_{m\in\widetilde{\mathcal{M}(\mathcal{D})}}$ *and* $z = \left(\frac{\log(\binom{p}{D_m})}{n}\right)_{m\in\widetilde{\mathcal{M}(\mathcal{D})}}$. *Let us define* $\theta = \left(C_1, C_2, C_3\right)^T$ *the vector of unknown parameters,* $W = \left(x|z|\mathbb{1}_d\right)$ *matrix of size* $\#\{m \in \widetilde{\mathcal{M}(\mathcal{D})}\} \times 3$, *A the diagonal matrix of size* 3×3 *with coefficients* $(1, 1, -1)$ *in the diagonal,* $b_0 = (0, 0, 0)^T$, $D = 2W^TW$, $u = 2W^Ty$ *and J the following function on* $\theta \in \mathbb{R}^3$*:*

$$J(\theta) = \frac{1}{2}\theta^T D\theta - u^T\theta \tag{5.15}$$

Then, minimizing 5.14 with respect to (C_1, C_2, C_3) *is equivalent to solve the problem of minimization under constraints with respect to* θ*:*

$$\underset{\substack{\theta\in\mathbb{R}^3 \\ s.t.\ A^T\theta\geq b_0}}{\arg\min} \left\{J(\theta)\right\} \tag{5.16}$$

Moreover, the solution of the problem of minimization under constraints (5.16) exists and is unique.

Proof of Proposition 5.1 can be found in Appendix 5.9.

5.4.2.2 Modification of the R function *DDSE* of the R package *Capushe*

In the following, we detail the modifications of the algorithm (2), step by step.

Step 1:
The main features on the model collection used for the slope estimation algorithm in the R function *DDSE* is the monotony of the $-\gamma_n\left(\hat{\beta}_m\right)$ values with respect to the $\text{pen}_{\text{shape}}(m)$ ones Thus, a preliminary implemented stage is to stop the collection as soon as one of these two properties is no longer verified. With our penalty (5.9), the function of the $\frac{\log\left(\binom{p}{D_m}\right)}{n}$ values with respect to the D_m ones decreases from $\lfloor\frac{p-1}{2}\rfloor + 1$. Removing all the models from this value allows us to recover one of the characteristics required for the *DDSE* algorithm. However, the risk is eliminating the large models corresponding to the expected affine behavior. Therefore, we propose to consider both possibilities in our algorithm: either stop the collection at $D_m = \lfloor\frac{p-1}{2}\rfloor + 1$, or conserve all the models.

Step 2:
As for the R function *DDSE*, two challenges have to be considered: select the triplets $\left(\frac{D_m}{n}, \frac{\log\left(\binom{p}{D_m}\right)}{n}, -\gamma_n\left(\hat{\beta}_m\right)\right)$ with large model complexities corresponding to the expected theoretical affine behavior; obtain a suitable estimation of the plane slope parameters. Instead

of searching for the best line passing through to a point cloud, the best affine trend in two dimensions is researching, giving parameters of the plane, especially the slope coefficients denoted $\Big(C_1(\sigma^2), C_2(\sigma^2)\Big)$. To implement the constrained minimization problem (5.16), we use the R function *solve.QP* of the R package *quadprog* minimizing $J(\theta) = \frac{1}{2}\theta^T D\theta - u^T\theta$ under the constraint $A^T\theta \geq b_0$. It is based on the dual method of Goldfarb and Idnani [Goldfarb and Idnani, 1983]. Briefly, from the unconstrained minimum as input, the principle is to find a violated constraint and update the solution by adding this constraint in the minimization; this is performed by using Cholesky and QR factorization to solve corresponding equations. These slope coefficients $\Big(C_1(\sigma^2), C_2(\sigma^2)\Big)$ give a available penalty function and the selected model can be computed. Then, as in the R function *DDSE*, our algorithm perform this coefficients estimation $\Big(C_1(\sigma^2), C_2(\sigma^2)\Big)$ of the affine behavior by removing one by one the model with the smallest dimension, until there are only three pairs of values $\left(\frac{D_m}{n}, \frac{\log\left(\binom{p}{D_m}\right)}{n}, -\gamma_n\left(\hat{\beta}_m\right)\right)$ (to estimate a plane, that is to say the two slope coefficients and the intercept, three pairs of values are required). At the end of this step, the function of the selected models, obtained by successively removing the model with the smallest dimension, with respect to the number of remaining couples of values is available.

Step 3:
This step corresponds to the first challenge of the slope estimation algorithm (see Subsection 5.3.2). Our implementation is almost identical to the R function *DDSE*. The selection of the plateau of the function of the selected models with respect to the number of remaining couples of values is identical. The only difference is that we consider two ways to get the final $\Big(\widehat{C_1}(\sigma^2), \widehat{C_2}(\sigma^2)\Big)$: the first one is by using the triplets $\left(\frac{D_m}{n}, \frac{\log\left(\binom{p}{D_m}\right)}{n}, -\gamma_n\left(\hat{\beta}_m\right)\right)$ from the beginning on the selected plateau until the end of the collection; the second one is slightly different by using only the triplets of the selected plateau.
Then, the final $\widehat{m}\Big(\widehat{C_1}(\sigma^2), \widehat{C_2}(\sigma^2)\Big)$ is calculated on the entire model collection by using (5.9) with $\Big(C_1(\sigma^2), C_2(\sigma^2)\Big) = \Big(\widehat{C_1}(\sigma^2), \widehat{C_2}(\sigma^2)\Big)$.

Here is the detail of our algorithm. The presentation of the algorithm includes R language. The notations data\$pen1, data\$pen2 and data\$LS stand for respectively the vector of the $\frac{D_m}{n}$ values, the $\frac{\log\left(\binom{p}{D_m}\right)}{n}$ ones and the vector of the least-squares values $\gamma_n\left(\hat{\beta}_m\right)$ along the model collection. The color blue indicates changes from the previous algorithm 2.

Algorithm 3: The affine trend in two dimensions coefficients estimation algorithm [1]

Data: data; pct; stop_researching; only_plateau

Result: $\widehat{m}_{\widehat{C_1},\widehat{C_2}}$; $\widehat{C_1}$; $\widehat{C_2}$

Step 1: On the model collection

```
data = data[-which(is.na(data), arr.ind = TRUE)[, 1], ] ;
data = data[order(data$pen1,data$pen2, data$LS), ] ;

plength = length(data$pen1);
for i in plength : 2 do
    if data$pen1[i] == data$pen1[i - 1] then
        data = data[-i, ];
    end
end

stop_increasing = which(diff(data$pen2) <0)[1] ;
# stop the collection as soon as the penalty shape decreases ;
if stop_researching then
    data = data[1:stop_increasing, ]
end
```

Algorithm 3: The affine trend in two dimensions coefficients estimation algorithm [2]

Step 2: Estimation of all the slopes

```
couplepen1 = data$pen1;
couplepen2 = data$pen2;
plength = length(couplepen1)-2;
couplecontrast = -data$LS;
C1 = numeric(plength);
C2 = numeric(plength);
mhat = numeric(plength);

Amat = matrix(c(1,0,0,0,1,0,0,0,-1),nrow = 3, ncol = 3);
bvec = c(0,0,0);

for p in 1 : plength do
    # constrained minimization method to estimate the slope coefficients;
    # estimation of the slope coefficients and the intercept of the best affine trend in
      two dimensions passing through the points cloud;
    # only the slope coefficients are conserved;

    Wmat = matrix(c(couplepen1,couplepen2,rep(1,length(couplepen1))),nrow =
      length(couplepen1), ncol = 3);
    Dmat = 2*t(Wmat)*Wmat;
    dvec = 2*(t(Wmat)*(couplecontrast));
    ResSolve.QP = solve.QP(Dmat,dvec,Amat,bvec,meq=0,factorized=FALSE);
    C1[p] = ResSolve.QP$solution[1];
    C2[p] = ResSolve.QP$solution[2];
    # at each step, the couple with the smallest dimension value is removed;
    couplepen1 = couplepen1[-1] ;
    couplepen2 = couplepen2[-1] ;
    couplecontrast = couplecontrast[-1];
    mhat[p] = which.min(data$LS + 2.(C1[p] data$pen1 + C2[p] data$pen2 ))
end
```

Algorithm 3: The affine trend in two dimensions coefficients estimation algorithm [2]

Step 3: Locating the affine behavior

```
Pi = c(1);
N = c();
number = 1;
for p in 2 : plength do
    if mhat[p] != mhat[p - 1] then
        Pi = c(Pi, p) ;
        N = c(N, number) ;
        number = 0;
    end
    number = number + 1
end
N = c(N, number);
Pi = c(Pi, p);
Imax = length(N);
while N[Imax] < (pct*sum(N)) do
    Imax = Imax - 1 ;
end
if only_plateau then
    interval_affine_behavior = Pi[Imax]:(Pi[Imax + 1] - 1)
end
else
    interval_affine_behavior = Pi[Imax]:(plength+2)
end
Wmat =
 matrix(c(data$pen1[interval_affine_behavior],data$pen2[interval_affine_behavior
          rep(1,length(data$pen1[interval_affine_behavior]))),
          nrow = length(data$pen1[interval_affine_behavior]), ncol = 3);
Dmat = 2*t(Wmat)*Wmat;
dvec = 2*(t(Wmat)*(-data$LS[interval_affine_behavior]));
ResSolve.QP = solve.QP(Dmat,dvec,Amat,bvec,meq=0,factorized=FALSE);
```

$\widehat{C_1}$ = ResSolve.QP$solution[1];

$\widehat{C_2}$ = ResSolve.QP$solution[2];

$\widehat{m}_{\widehat{C_1},\widehat{C_2}}$ = which.min$\Big($data$LS + 2.$($\widehat{C_1}$ data$pen1 + $\widehat{C_2}$ data$pen2)$\Big)$

Rewriting the hyperparameters calibration problem as a constrained minimization problem is advantageous because the slope coefficients are now positive. However, we lose robustness since our procedure no longer contains robust linear regressions. Thus, the plateaus of the function of

the successive selected models with respect to the number of triplets $\left(\frac{D_m}{n}, \frac{\log(\binom{p}{D_m})}{n}, -\gamma_n\left(\hat{\beta}_m\right)\right)$ used for the linear regression coefficient estimation are significantly more numerous and smaller. Thus, our algorithm is very sensitive to the input parameter "*pct*".

5.5 Adding the randomness of the model collection in the slope estimations

We modify our algorithm (3) to take into account the randomness of the model collection.

5.5.1 Resampling procedure

An idea is to replicate the model collection construction to enrich the collection of available models. For that, we are inspired by the sampling strategies Bolasso [Bach, 2008] and Stability Selection [Meinshausen and Bühlmann, 2010] whose main idea is to vary the original dataset slightly. More precisely, we generate, from the initial dataset, R resamples: $\mathcal{D}^1 = (Y^1, X^1), \cdots, \mathcal{D}^R = (Y^R, X^R)$. Each resample is a sampling of the n original observations uniformly chosen with replacement. Then, for $r \in \{1, \cdots, R\}$, a model collection $\mathcal{M}_r(\mathcal{D}^r)$ is generated on each $\mathcal{D}^r$ independently of each other. To further support the random nature of the model collections creation and limit the resampling effects, we propose to add the Randomized Lasso principle, introduced in [Meinshausen and Bühlmann, 2010], to construct the random model collections. It consists in a random perturbation on each variable: instead of minimizing the regularized least-squares with respect to $\beta \in \mathbb{R}^p$, it is minimizing with respect to $\frac{\beta_j}{W_j}$ where $W_j \sim \mathcal{U}\big([\gamma, 1]\big)$, for $\gamma > 0$. Considering several model collections generated from different resamples allows to get robustness to some outliers potentially present in the original dataset.
At the end, a collection $\Big(\mathcal{M}_1(\mathcal{D}^1), \cdots, \mathcal{M}_R(\mathcal{D}^R)\Big)$ of model collections is available, given a list of data\$pen1, of data\$pen2, of data\$LS_1, and of data\$LS_2. Each element of a list corresponds respectively to the data\$pen1, the data\$pen2, the data\$LS_1 and the data\$LS_2 values obtained in one resample. The data\$LS_1 values from $\mathcal{D}^r$ correspond to the least-squares evaluated on $\mathcal{D}^r$, whereas concerning the data\$LS_2 values from $\mathcal{D}^r$, the $\left(\hat{\beta}_m\right)_{m \in \mathcal{M}_r(\mathcal{D}^r)}$ are computed on $\mathcal{D}^r$ but the least-squares are evaluated on the entire dataset $\mathcal{D}$. In the following, we propose to consider the randomness of the collection in our algorithm (Subsection 5.5.2).

5.5.2 Improvement of our algorithm 3 with a random model collection

The main objective is to generalize the R function *DDSE* of the R package *Capushe* to take into account the randomness of the model collection, in addition to calibrate $C_1(\sigma^2)$ and $C_2(\sigma^2)$

simultaneously with the positiveness constraints. We add the resampling process to our algorithm (3). In the following, we detail the modifications of the algorithm (3), step by step, starting with the list of data$pen1, of data$pen2, of data$LS_1, and data$LS_2.

Step 1:
We consider three different model collections to estimate the best affine trend in two dimensions. They are all constructed from $\Big(\mathcal{M}_1(\mathcal{D}^1), \cdots, \mathcal{M}_R(\mathcal{D}^R)\Big)$.

1. The model collection obtained by keeping only one model per dimension: the one minimizing the least-squares evaluated on each resample for the first one;
2. The model collection obtained by minimizing the least-squares evaluated on the entire dataset for the second one;
3. The model collection obtained by keeping all the obtained models: $\bigcup_{r=1}^{R} \mathcal{M}_r(\mathcal{D}^r)$.

As in our algorithm (3), we take into account the possibility to stop the collection as soon as the complexity decreases (starting at $D_m = \lfloor \frac{p-1}{2} \rfloor + 1$).

Step 2:
If we process the test H_0="no effect of the collection on the least-squares estimations" against H_1="there exists an effect of the collection on the least-squares estimations", the corresponding p-value is smaller than 10^{-16}. This too small value means that the collection model affects the least-squares values. Therefore, an *ANOVA* (for *Analysis of variance*) is processed in a preliminary step removing these effects from the model collection. More precisely, the ANOVA goal is to remove the variation between the different collections and the variation within each collection. It is based on the minimization of the quantity $|| - \gamma_n\Big(\hat{\beta}_m\Big) - (-\tilde{L}S)||_{2,n}^2$ where the component i of the vector $-\tilde{L}S$ is the least squares average of all models of the collection to which the i-th component of the vector $-\gamma_n\Big(\hat{\beta}_m\Big)$ belongs. The ANOVA step leads to new least-squares values which are used to estimate the pairs $\Big(C_1(\sigma^2), C_2(\sigma^2)\Big)$. After this preliminary step, the implementation of the coefficients estimations of the best affine trend in two dimensions is the same as in (3) except that when the considered model collection is $\bigcup_{r=1}^{R} \mathcal{M}_r(\mathcal{D}^r)$, where there are several models with the same dimension. In this last case, each step corresponds to coefficients estimations when all models with the smallest dimension of the remaining cloud of values are removed. Concerning the $\widehat{m}\Big(C_1(\sigma^2), C_2(\sigma^2)\Big)$ values, they are computed from all the data (with the data$LS_2 list).

Step 3:
This step is exactly the same as our algorithm (3).

Here is the detail of our algorithm. The presentation of the algorithm includes R language. The notations data\$pen1, data\$pen2, data\$LS_1 and data\$LS_2 stand for respectively the list of vectors of the $\frac{D_m}{n}$ values, the $\frac{\log(\binom{p}{D_m})}{n}$ ones, the list of vectors of the least-squares values $\gamma_n(\hat{\beta}_m)$ evaluated on each resample and the list of vectors of the least-squares values $\gamma_n(\hat{\beta}_m)$ evaluated on the entire dataset. The color blue indicates changes from the previous algorithm 3.

Algorithm 4: The affine trend in two dimensions coefficients estimation algorithm with the collection randomness [1]

Data: data; pct; stop_researching; LS_full_data; only_plateau; all_models

Result: $\widehat{m}_{\widehat{C_1},\widehat{C_2}}$; $\widehat{C_1}$; $\widehat{C_2}$

Step 1: On the model collection

```
number_shuffles = length(data$pen1);
data$number = unlist(sapply(1:number_shuffles, function(it)
 rep(it,length(data$pen1[[it]]))));
data$pen1 = unlist(sapply(1:number_shuffles, function(it) data$pen1[[it]]));
data$pen2 = unlist(sapply(1:number_shuffles, function(it) data$pen2[[it]]));
data$LS_2 = unlist(sapply(1:number_shuffles, function(it) data$LS_2[[it]]));

if LS_full_data then
  data$LS = data$LS_2 ;
end
else
  data$LS = unlist(sapply(1:number_shuffles, function(it) data$LS_1[[it]]));
end

data = data[-which(is.na(data), arr.ind = TRUE)[, 1], ] ;
data = data[order(data$pen1,data$pen2, data$LS), ] ;

if all_models == FALSE then
  # If several models in the collection have the same penalty shape value, keeping
    only the one with the smallest least-squares value ;
  plength = length(data$pen1);
  for i in plength : 2 do
    if data$pen1[i] == data$pen1[i - 1] then
      data = data[-i, ];
    end
  end
end

stop_increasing = which(diff(data$pen2) <0)[1] ;
if stop_researching then
  data = data[1:stop_increasing, ]
end
```

Algorithm 4: The affine trend in two dimensions coefficients estimation algorithm with the collection randomness[2]

Step 2: Estimation of all the slopes

```
couplepen1 = data$pen1;
couplepen2 = data$pen2;

if all_models then
    plength = length(unique(data$pen1))-2 ;
    # models are removed per groups of same dimension ;
end
else
    plength = length(couplepen1)-2
end
C1 = numeric(plength);
C2 = numeric(plength);
mhat = numeric(plength);
couplecontrast = rlm(-data$LS ~ as.factor(data$number) -1)$residuals ;
# effects of the resample number are removed ;

Amat = matrix(c(1,0,0,0,1,0,0,0,-1),nrow = 3, ncol = 3);
bvec = c(0,0,0);
for p in 1 : plength do
    Wmat = matrix(c(couplepen1,couplepen2,rep(1,length(couplepen1))),nrow =
      length(couplepen1), ncol = 3);
    Dmat = 2*t(Wmat)*Wmat;
    dvec = 2*(t(Wmat)*(couplecontrast));
    ResSolve.QP = solve.QP(Dmat,dvec,Amat,bvec,meq=0,factorized=FALSE);
    C1[p] = ResSolve.QP$solution[1];
    C2[p] = ResSolve.QP$solution[2];

    if all_models then
        # at each step, all the couples with the smallest dimension value are removed;
        couplepen1 = couplepen1[-which(couplepen1 == unique(couplepen1)[1])] ;
        couplepen2 = couplepen2[-which(couplepen1 == unique(couplepen1)[1])] ;
        couplecontrast = couplecontrast[-which(couplepen1 == unique(couplepen1)[1])];
    end
    else
        # at each step, the couple with the smallest dimension value is removed;
        couplepen1 = couplepen1[-1] ;
        couplepen2 = couplepen2[-1] ;
        couplecontrast = couplecontrast[-1];
    end
    # the selected model depends on the least-squares computed on the entire dataset ;
    mhat[p] = which.min(data$LS_2 + 2.(C1[p] data$pen1 + C2[p] data$pen2 ))
end
```

Algorithm 4: The affine trend in two dimensions coefficients estimation algorithm with the collection randomness[3]

```
Step 3: Locating the affine behavior
Pi = c(1);
N = c();
number = 1;

for p in 2 : plength do
    if mhat[p] != mhat[p - 1] then
        Pi = c(Pi, p) ;
        N = c(N, number) ;
        number = 0;
    end
    number = number + 1
end

N = c(N, number);
Pi = c(Pi, p);
Imax = length(N);
while N[Imax] < (pct*sum(N)) do
    Imax = Imax - 1 ;
end

if all_models then
    if only_plateau then
        interval_affine_behavior =
          (which(data$pen1>=unique(data$pen1)[Pi[Imax]])[1]):length(data$pen1)
    end
    else
        interval_affine_behavior =
          (which(data$pen1==unique(data$pen1)[Pi[Imax]])[1]):
                        ((which(data$pen1==unique(data$pen1)[Pi[Imax+1]])[1])-1)
    end
end
else
    if only_plateau then
        interval_affine_behavior = Pi[Imax]:(Pi[Imax + 1] - 1)
    end
    else
        interval_affine_behavior = Pi[Imax]:(plength+2)
    end
end
```

Algorithm 4: The affine trend in two dimensions coefficients estimation algorithm with the collection randomness[3]

End of step 3: Locating the affine behavior

```
Wmat =
 matrix(c(data$pen1[interval_affine_behavior],data$pen2[interval_affine_behavior],;
          rep(1,length(data$pen1[interval_affine_behavior]))), ;
          nrow = length(data$pen1[interval_affine_behavior]), ncol = 3);
Dmat = 2*t(Wmat)*Wmat;
dvec = 2*(t(Wmat)*(-data$LS[interval_affine_behavior]));
ResSolve.QP = solve.QP(Dmat,dvec,Amat,bvec,meq=0,factorized=FALSE);
```

$\widehat{C_1}$ = ResSolve.QP$solution[1];

$\widehat{C_2}$ = ResSolve.QP$solution[2];

$\widehat{m}_{\widehat{C_1},\widehat{C_2}}$ = which.min$\Big($data$LS + 2.$\big($$\widehat{C_1}$ data$pen1 + $\widehat{C_2}$ data$pen2 $\big)\Big)$

5.6 Numerical evaluations of the methods

5.6.1 Simulation framework

Data from a Gaussian graphical model

The simulation protocol is the same as the one introduced in Chapter 3. We choose the Gaussian graphical model to generate the variables $\Big(Y, X_1, \cdots, X_p\Big)$ to obtain Gaussian data, dependency structures between variables and the most favorable case where assumptions on our model broadly hold. A sample of size n is generated from a $(p+1)$ multivariate centered Gaussian distribution with covariance matrix Σ, where the dependency structure is encoded in the precision matrix Σ^{-1}: if its coefficient $(j, j^{'})$ is non-zero, it means that the variables X_j and $X_{j'}$ interact when all other variables are given. We refer to [Lauritzen, 1996] and [Córdoba et al., 2019] for more information. The response variable Y is chosen as a column of the p multivariate centered Gaussian. Then, Y is removed from the matrix X so that X is a matrix of size $n \times p$. To evaluate our algorithms in a fair way, we consider two types of dependency:

- Cluster design: the precision matrix is simulated as a block diagonal matrix with B blocks of equal size where B divides $(p+1)$. The response variable Y is defined as the first variable.
- Scale-free design: a few variables have a lot of neighbors while all the others have a few ones. The response variable Y is defined as the variable having the highest number of neighbors. This simulation design is named scale-free-max in the sequel.

Data parameters

Data parameters are close to those in the comparison of existing variable selection procedures

using regularization paths (see Chapter 3). We set $n = 150$, $p = 199$ and 40 samples of size $2n$ are generated to create a training set of size n used for the estimations (model collection generations, slope parameters calibrations, model selection) and a validation set of size n used for the mean squared errors evaluations. For this chapter, 40 samples are generated. We use the R function *huge.generator* from the R package *huge* (version 1.3.4.1) with block number B equals 5 and probability of connection within a component setting to the default value 0.3 for the cluster design.
When resampling is performed, 50 resamples are generated from the original data set. Concerning the Randomized Lasso principle, weights W_j are drawn from a uniform distribution between 0.7 and 1.

About the model collection
Because of the high-dimensional context and the presence of unknown strongly correlations between variables, we use the Elastic-Net penalized least-squares minimization [Zou and Hastie, 2005]:

$$\underset{\beta \in \mathbb{R}^p}{\arg\min}\Big\{\gamma_n(\beta) + \lambda\Big((1-\alpha)||\beta||_2^2 + \alpha|\beta|_1\Big)\Big\} \tag{5.17}$$

where $\lambda > 0$ and $\alpha \in [0,1]$, to generate the regularization path $\Big[\lambda \mapsto \hat{\beta}_\lambda\Big]$. This last one corresponds to a set of variable subsets. Each subset is associated with a model m, which corresponds to the subspace of $\mathbb{R}^p$ spanned by the variables of the subset.
Since the R function *DDSE* assumes the monotony of the $-\gamma_n\Big(\hat{\beta}_m\Big)$ values with respect to the $\text{pen}_{\text{shape}}(m)$ ones and the monotony of the $\text{pen}_{\text{shape}}(m)$ values with respect to the complexity of models, we propose to use the LARS algorithm to generate model collections through the Elastic-Net penalized least-squares minimization (5.17). Briefly speaking, LARS leads to almost nested models given a decreasing value of the $-\gamma_n\Big(\hat{\beta}_m\Big)$ with respect to the dimension of the model. Thus, models in the generated collection have to be almost well-ordered. This choice is consistent with the results of the chapter 3: the combination E-Net regularization and lars algorithm provides the best results whatever the evaluated metric. We use the function *enet* of the R package *elasticnet* (version 1.3). The maximal number of steps to define the grid size is the default value $50 \times \min(p, n-1)$. We set $\alpha = 0.5$ for the Elastic-Net regularization.

About the model selection
LinSelect is implemented in the function *tuneLasso* of the R package *LINselect* (version 1.1.3). The parameters are set to the default values, especially the multiplicative coefficient of the penalty function fixed to 1.1. The slope estimation in the data-driven penalties is calculated by using the function *DDSE* of the R package *capushe* (version 1.1.1). For the *rlm* application, the option "*psi.bisquare*" is chosen corresponding to the bi-squared function of Tukey for the M-estimator. Lastly, the minimum percentage of points required for the plateau selection is setting to 0.05, both for the line slope and the plane slope coefficients estimations. For the constrained minimization, the *solve.QP* R function of the R package *quadprog* (version 1.5.8) is used with all parameters fixed by default.

5.6.2 Analyses of the output characteristics of the proposed algorithms

Most of the Figures used in this subsection can be found in Appendix 5.10. This subsection stands for study characteristics of the data, a reminder of the methods submitted for comparison and analyses of the outputs of the proposed slope estimation algorithms.

5.6.2.1 Characteristics of our data

Table 5.1: Active variable number

$\lvert\beta^*\rvert$	cluster	scale_free_max
median	12	30
mean	11.9	30.9
(sd)	*(2.75)*	*(9.56)*

Table 5.1 summarizes the size of the support of β^* for both cluster and scale-free-max designs. The sizes do not take extreme values, the sparsity hypothesis is respected. In addition to evaluate our algorithms on two different dependency structures, many support sizes are taken into account in the simulation study, allowing to fully investigate the method performances under various support sizes: means and medians are close to 12 and 30 respectively for cluster and scale-free-max, and the standard deviations are high, especially for scale-free-max (9.56).

5.6.2.2 Evaluated methods

Our algorithms aim to improve slope heuristics methods in terms of prediction performances. We consider the data-driven penalties via the slope heuristics strategies and LinSelect penalty function in this simulation study. The reason is that these are the only penalties adapted to our framework (5.1) with a high-dimensional context, an unknown variance and a large model collection and leading to non-asymptotic guarantees on the predictive risk (see Chapter 3).

We propose the following naive approach to validate the necessity of using data-driven penalties containing two hyperparameters. It consists in the exploration on $\mathcal{C}$, a fixed grid of couples (C_1, C_2). For each $(C_1, C_2) \in \mathcal{C}$, the selection model $\hat{m}_{(C_1,C_2)}$ is:

$$\hat{m}_{(C_1,C_2)} = \underset{m\in\mathcal{M}(\mathcal{D})}{\arg\min}\left\{\gamma_n(\hat{\beta}_m) + 2\left(C_1\frac{D_m}{n} + C_2\frac{\log(\binom{p}{D_m})}{n}\right)\right\}$$

The predictive risk values are evaluated for each $\hat{m}_{(C_1,C_2)}$ by the mean squared errors (MSE) defined by: $\forall m \in \mathcal{M}$:

$$\mathrm{MSE}(m) = \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\Big(\tilde{y}_i - \sum_{j=1}^{p}\tilde{x}_{ij}\hat{\beta}_{m,j})_i\Big)^2 \tag{5.18}$$

where $\tilde{\mathcal{D}} = (\tilde{X}, \tilde{Y})$ is a data set of size $\tilde{n}$ independent of $\mathcal{D}$ and identically distributed. Lastly, the final model is:

$$\hat{m} = \underset{(C_1,C_2)\in\mathcal{C}}{\arg\min} \left\{ \text{MSE}\left(\hat{m}_{(C_1,C_2)}\right) \right\}$$

To add the random aspect of the collection in the naive approach, we propose to solve:

$$\hat{m}_{(C_1,C_2)} = \underset{m\in \bigcup_{r=1}^{R} \mathcal{M}_r(\mathcal{D}^r)}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_{m_j}\right)^2 + 2\left(C_1\frac{D_m}{n} + C_2\frac{\log(\binom{p}{D_m})}{n}\right)\right\},$$

where we consider in the minimization $\bigcup_{r=1}^{R} \mathcal{M}_r(\mathcal{D}^r)$, for each $(C_1, C_2) \in \mathcal{C}$. Note that $\hat{\beta}_m$ is computed on $\mathcal{D}^r$ for $m \in \mathcal{M}_r(\mathcal{D}^r)$ while the least-squares are evaluated on the entire dataset $\mathcal{D}$. Then, the rest of the naive approach with a random model collection is identical to the previous one.
This naive approach, expensive computationally, can be considered as a benchmark to compare our methods calibrating the couple $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ automatically when the randomness of the collection is taken into account. As all constants are tested through a grid, this naive approach can be seen as the best we can do in our context, with the penalty (5.9) and a data-dependent collection. As all constants are tested through a grid, this naive approach can be seen as the best we can do in our context. To apply the naive approach, we use the grid $\mathcal{C} = [0, 10] \times [0, 10]$.

Finally, the methods that we evaluate and compare are listed below. They are accompanied by their diminutive used thereafter:

- *Capus_ C2_ 0*: is the existing slope estimation algorithm (Subsection 5.3) with $C_2(\sigma^2) = 0$ and using the R function *DDSE* of the R package *Capushe*
- *Capus_ 2.5*: is the existing slope estimation algorithm (Subsection 5.3) with $\frac{C_1(\sigma^2)}{C_2(\sigma^2)} = 2.5$ and using the R function *DDSE* of the R package *Capushe*
- *LinSelect*: is the LinSelect penalty defined in (2.21)
- *grid*: is the naive approach
- *2D_ Stop_ End*: is the generalization of slope estimation coefficients by processing a constrained minimization with options *stop_ researching == TRUE* and *only_ plateau == TRUE* (stop the collection at $D_m = \lfloor \frac{p-1}{2} \rfloor + 1$ and triplets used to estimate hyperparameters are those from the beginning on the selected plateau until the end of the collection)
- *2D_ Stop_ Plat*: is the generalization of slope estimation coefficients by processing a constrained minimization with options *stop_ researching == TRUE* and *only_ plateau == FALSE* (stop the collection at $D_m = \lfloor \frac{p-1}{2} \rfloor + 1$ and triplets used to estimate hyperparameters are those of the selected plateau)

- *2D_Full_End*: is the generalization of slope estimation coefficients by processing a constrained minimization with options *stop_researching* == *FALSE* and *only_plateau* == *TRUE* (conserve all the models and triplets used to estimate hyperparameters are those from the beginning on the selected plateau until the end of the collection)

- *2D_Full_Plat*: is the generalization of coefficients by processing a constrained minimization with options *stop_researching* == *FALSE* and *only_plateau* == *FALSE* (conserve all the models and triplets used to estimate hyperparameters are those of the selected plateau).

These 8 methods are implemented from 4 different model collection types:

- *path*: where $\mathcal{M}(\mathcal{D})$ is generated by the LARS algorithm with the Elastic-Net penalized least-squares minimization (5.17) on the entire dataset $\mathcal{D}$

- *path_Bolasso_LS_each_resample*: generated from $\left(\mathcal{M}_1(\mathcal{D}^1), \cdots, \mathcal{M}_R(\mathcal{D}^R)\right)$ by keeping only one model per dimension: minimizing the least-squares evaluated on each $\left(\mathcal{D}^r\right)_{r\in\{1,\cdots,R\}}$ (when *LS_full_data* == *FALSE*)

- *path_Bolasso_LS_fulldata*: generated from $\left(\mathcal{M}_1(\mathcal{D}^1), \cdots, \mathcal{M}_R(\mathcal{D}^R)\right)$ by keeping only one model per dimension: minimizing the least-squares evaluated on the entire dataset (when *LS_full_data* == *TRUE*)

- *path_Bolasso_all_models*: the one consisting in keeping all the obtained models: $\bigcup_{r=1}^{R} \mathcal{M}_r(\mathcal{D}^r)$ (when *all_models* == *TRUE*).

Hence, 32 methods are compared, corresponding to all combinations of the previous model collection types and the studied penalization functions.

Note that as we don't want to modify the code of the R function *DDSE* of the R package *Capushe*, applying *Capus_C2_0* and *Capus_2.5* on the second model collection is identical to apply them on the fourth one. Indeed, the first step of the R function *DDSE* consists in reducing the available collection to a single model per dimension by keeping the model with the smallest least-squares value, so *Capus_C2_0* (respectively *Capus_2.5*) on *path_Bolasso_LS_each_resample* is identical to *Capus_C2_0* (respectively *Capus_2.5*) on *path_Bolasso_LS_fulldata*.

5.6.2.3 Analysis of the proposed slope coefficients estimation algorithms

This part is devoted to analyzing the proposed slope coefficient estimation algorithms' characteristics. Firstly, the resampling addition in the model collection is evaluated; secondly,

the step of hyperparameters calibration $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ is studied, in particular, outputs $\left(\widehat{C_1}(\sigma^2), \widehat{C_2}(\sigma^2)\right)$ are compared with the $\widehat{\kappa(\sigma^2)}$ ones; thirdly, the ability to detect the expected theoretical affine behavior is analyzed.

resampling to generate the model collection: In this paragraph, we propose to analyze the different types of model collections considered in this simulation study in both cluster and scale-free-max designs.

Figures 5.6 plots the values of $-\gamma_n(\hat{\beta}_m)$ with respect to D_m for the 8 first model collections $\left(\mathcal{M}_1(\mathcal{D}^1), \cdots, \mathcal{M}_8(\mathcal{D}^8)\right)$. We observe similar shapes on the 8 curves suggesting that the collections of models are all relevant on the resamples. Moreover, a higher variability between curves is obtained on the cluster design leading to more instability of the Elastic-Net criterion minimization (5.17) in this dependency structure.

Figures 5.7 plots the values of $-\gamma_n(\hat{\beta}_m)$ with respect to D_m of the four model collections considered at the output of step 1 of the algorithms.
For the first three (*path*, *path_Bolasso_LS_each_resample*, and *path_Bolasso_LS_fulldata*), there is only one model per dimension, and so correspond to only one curve (respectively red, green and cyan) in Figures 5.7. Concerning the last one (*path_Bolasso_all_models*), the model collection contains all the models of collections from resamples and so corresponds to the 50 blue curves in Figures 5.7. The correlation structure does not seem to impact the generation of model collections. As expected, the *path_Bolasso_LS_each_resample* curves encapsulate the curves of paths from each of the $\left(\mathcal{M}_r(\mathcal{D}^r)\right)_{r\in\{1,\cdots,R\}}$. The *path* curve is often below all the curves obtained from the resamples. One explanation may be that there is less variability within the resamples rather than within the original dataset, leading to estimators $\hat{\beta}_m$ with better prediction performances and so to smaller least-squares values along the path. Concerning the *path_Bolasso_LS_fulldata* curves, the least-squares values are deteriorating for large dimensions. Hence, the monotonicity of the least squares with respect to the dimension can not be used anymore. However, this phenomenon could avoid the selection of a too large model. Moreover, curves from *path_Bolasso_LS_fulldata* are the closest one to the *path* ones for small dimensions.

The step 2: estimation of C1 and C2 In this paragraph, we propose to analyze the estimated hyperparameters $\left(\widehat{C_1}(\sigma^2), \widehat{C_2}(\sigma^2)\right)$ in comparison with $\widehat{\kappa(\sigma^2)}$. This part is devoted to evaluating the second challenge of the slope coefficients estimation algorithm: how to obtain a proper estimation of the slope given a reasonable hyperparameter calibration?

A first step emphasizing that our proposed procedures are different from the state of the art is, on the one hand, to count the number of time that the hyperparameter $C_2(\sigma^2)$ is

estimated to 0 (case of *Capus_C2_0*), and on the other hand, to study the ratio $\frac{\widehat{C_1}(\sigma^2)}{\widehat{C_2}(\sigma^2)}$ compared to 2.5 (case of *Capus_2.5*). Tables 5.2 and 5.3 summarize the $\widehat{C_2}(\sigma^2)$ nulls accounts. In our framework, as the used model collection are random and data-dependent, a non-zero value of $\widehat{C_2}(\sigma^2)$ is expected For both designs, the smallest values are obtained with *2D_Full_End* For cluster, the model collection generation by resampling strategies given the smallest values are *path_Bolasso_LS_fulldata* and *path_Bolasso_all_models* whereas it is *path_Bolasso_LS_each_resample* for scale-free-max Thus, our new procedures seem to differ from *Capus_C2_0*, especially when resampling is considered.

Table 5.2: cluster

	2D_Stop_End	2D_Stop_Plat	2D_Full_End	2D_Full_Plat
path	17	21	12	11
path_Bolasso_LS _on_each_resample	11	11	6	13
path_Bolasso_LS_fulldata	3	8	0	4
path_Bolasso_all_models	3	8	0	4

Table 5.3: scale-free-max

	2D_Stop_End	2D_Stop_Plat	2D_Full_End	2D_Full_Plat
path	14	15	13	14
path_Bolasso_LS _on_each_resample	4	5	5	10
path_Bolasso_LS_fulldata	10	12	0	11
path_Bolasso_all_models	10	12	0	11

Figures 5.8 and 5.9 are devoted to the $\frac{\widehat{C_1}(\sigma^2)}{\widehat{C_2}(\sigma^2)}$ ratio values for respectively, the cluster and the scale-free-max designs. Here, the potential infinite values of the ratios are removed. The first important observation is that the ratio is never equal or close to 2.5 Moreover, the ratios are never constant. Thus, our new algorithms seem to differ from *Capus_2.5*. The ideal ratio values are close to those obtained by the *grid* approaches, considered like the benchmarks. Excepted with *path* on the scale-free-max design where the ratio median is close to 1, all the others are close to 0. Hence, the values of $\widehat{C_1}(\sigma^2)$ are small compared to those of $\widehat{C_2}(\sigma^2)$. Contrary to *Capus_C2_0* and *Capus_2.5* where the values of $\left(\frac{D_m}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ have more weights in the penalization, these are the $\left(\frac{\log\left(\binom{p}{D_m}\right)}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ in both naive approaches and our penalization functions. For both designs, stopping the collection at $D_m = \lfloor\frac{p-1}{2}\rfloor + 1$ (*2D_Stop_End* and *2D_Stop_Plat*) leads to ratio values the closest to the *grid* approaches ones. When all the models of the collection are conserved (*2D_Full_End* and *2D_Full_Plat*), the ratio values are

close to 1 meaning that the $\left(\frac{D_m}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ and the $\left(\frac{\log\left(\binom{p}{D_m}\right)}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ values are assigned to the same weight in the penalty functions.

Lastly, Figures 5.10 and 5.11 are devoted to the study of both $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ separately. For both cluster and scale-free-max designs, and for both $C_1(\sigma^2)$ and $C_2(\sigma^2)$ estimations, stopping the collection at $D_m = \lfloor\frac{p-1}{2}\rfloor + 1$ (*2D_Stop_End* and *2D_Stop_Plat*) leads to values closer to the *grid* approaches ones than by conserving all the models of the collection (*2D_Full_End* and *2D_Full_Plat*). More precisely, when triplets used to estimate hyperparameters are those from the beginning on the selected plateau until the end of the collection (*2D_Stop_End*), the method is closer to the naive than when triplets used to estimate hyperparameters are those of the selected plateau (*2D_Stop_Plat*). Except for some extreme values, $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ are contained between 0 and 2 and they are lower with *path_Bolasso_all_models*. Lastly, values from *Capus_C2_0* and *Capus_2.5* are the farthest away proving that $\widehat{C_1}(\sigma^2)$ values are far from the $\widehat{\kappa(\sigma^2)}$ ones, as well as $\widehat{C_2}(\sigma^2)$. Moreover, some values are negative, showing the failure of the use of the *Capus_C2_0* and *Capus_2.5* data-driven penalties. Finally, estimated hyperparameters from our algorithms are higher than from the *Capus_C2_0* and *Capus_2.5* penalties. Thus, our penalty functions take higher values leading to more conservative procedures than the existing ones.

To conclude this part, adding a minimization under constraints procedure to estimate both hyperparameters leads to values of $\widehat{C_1}(\sigma^2)$, $\widehat{C_2}(\sigma^2)$ and $\frac{\widehat{C_1}(\sigma^2)}{\widehat{C_2}(\sigma^2)}$ closer to the *grid* approaches ones, which can be considered as benchmarks, compared to the existing *Capus_C2_0* and *Capus_2.5*. The *2D_Stop_End* and *2D_Stop_Plat* strategies lead to values closer to the *grid* ones than the *2D_Full_End* and *2D_Full_Plat* strategies. It could be due to the noise leading to some estimation errors.

The step 3: expected theoretical affine behavior: Instead of looking for a line from a curve in one dimension (see Figure 5.1), a plane is to be looked for from a curve in two dimensions in our algorithms. In Figure 5.2 are plotted an example of the $-\gamma_n\left(\hat{\beta}_m\right)_{m\in\mathcal{M}(\mathcal{D})}$ values as a function of the $\left(\frac{D_m}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ and $\left(\frac{\log(\binom{p}{D_m})}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ ones for each cluster and scale-free-max design. Planes seem to be visible for model dimensions large enough.

In this paragraph, we propose to study the presence of the expected affine behavior and the ability of each studied method to detect it. This part is devoted to evaluate the first challenge of the slope coefficients estimation algorithm: how to select the couples $\left(\mathrm{pen}_{\mathrm{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ with large $\mathrm{pen}_{\mathrm{shape}}(m)$ values and satisfying the expected theoretical affine relation?

In Figures 5.12- 5.15, the least-squares values are plotted as a function of minimal penalty ones obtained by the *Capus_C2_0*, *Capus_2.5*, *2D_Stop_End*, *2D_Stop_Plat*, *2D_Full_End*,

[1]The figure is extracted from [Baudry et al., 2012]

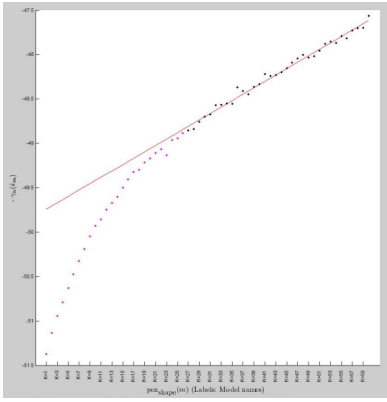
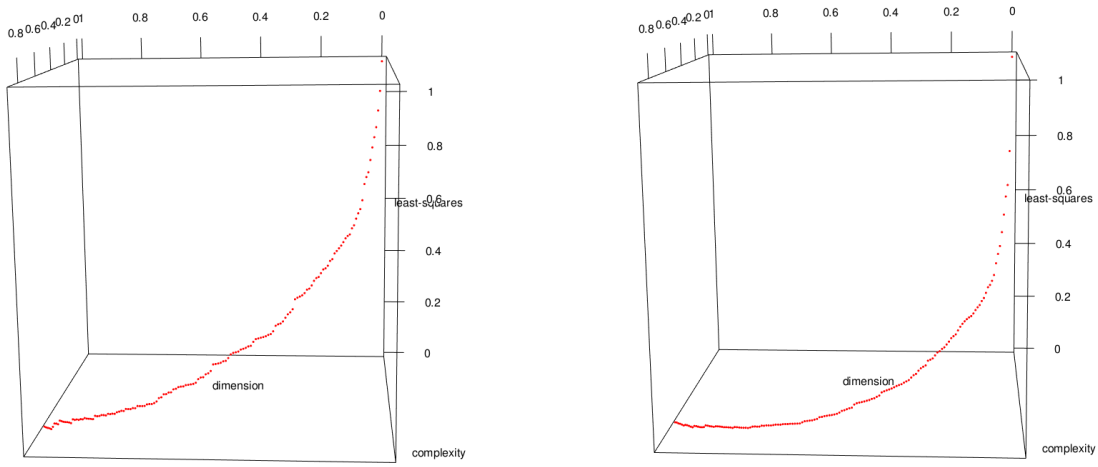


Figure 5.1: An output example of the R function *DDSE* of the R package *Capushe* illustrating the affine behavior between the $-\gamma_n\left(\hat{\beta}_m\right)_{m\in\mathcal{M}(\mathcal{D})}$ values and a penalty shape ones with only one unknown hyperparameter [1].



cluster scale-free-max

Figure 5.2: Examples of the $-\gamma_n\left(\hat{\beta}_m\right)_{m\in\mathcal{M}(\mathcal{D})}$ values as a function of the $\left(\frac{D_m}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ and $\left(\frac{\log(\binom{p}{D_m})}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ ones.

2D_Full_Plat methods from *path* or *path_Bolasso_LS_each_resample* model collection and for respectively cluster and scale-free-max designs. For each of them, the expected theoretical affine behavior is visible for large models which justifies the slope heuristics applications. It can be considered unexpected since many approximations have been made from the precise shape of the minimal penalty (5.5) to the penalty used for practical consideration (5.7). For instance, the bias becomes very small when D_m increases, but we suppose that it vanishes; the theory is based on a fixed and non-random collection, but we use the LARS algorithm leading to an approximation of the exact solution of the Elastic-Net criterion minimization which is itself a

replacement of the ideal minimization using the ℓ_0-norm.
Moreover, whatever the designs, the affine behaviors are visible on more models when only one hyperparameter is considered (*Capus_ C2_ 0* and *Capus_ 2.5*) than when two ones are considered (our proposed penalty functions). The affine behavior detected are unsurprisingly just before $D_m = \frac{p}{2}$ when model collection is stopping at $D_m = \lfloor \frac{p-1}{2} \rfloor + 1$ (*2D_ Stop_ End* and *2D_ Stop_ Plat*); and at the end of the collection when all models on the collection are conserving (*2D_ Full_ End* and *2D_ Full_ Plat*) for both resampling addition or not. These observations are consistent with those that can be observed on Figures 5.16 and 5.17: firstly, whatever the design, the number of models used to calibrate hyperparameters $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ is always significantly larger than 3 meaning than an affine relationship is detected between the least squares values and the penalty shapes ones; secondly, values are higher for *Capus_ C2_ 0* and *Capus_ 2.5* than our proposed penalty functions; thirdly, values are higher when all models on the collection are conserving (*2D_ Full_ End* and *2D_ Full_ Plat*) than when model collection is stopping at $D_m = \lfloor \frac{p-1}{2} \rfloor + 1$ (*2D_ Stop_ End* and *2D_ Stop_ Plat*). The second observation is explained since if the affine behavior is visible on more models, more models will be used to estimate slope parameters. The third observation is explained since expected theoretical affine behavior concerns large models of which a certain number are probably withdrawn on the two types of model collections *2D_ Stop_ End* and *2D_ Stop_ Plat*. Likewise, these characteristics are observed on Figures from 5.18 to 5.21 where are plotted the function of the successive selected models with respect to the number of couples $\left(\mathrm{pen}_{\mathrm{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ used for the affine regression coefficients estimation: whatever the design, plateaus are less visible when model collection is stopping at $D_m = \lfloor \frac{p-1}{2} \rfloor + 1$ (*2D_ Stop_ End* and *2D_ Stop_ Plat*) than when all models on the collection are conserving (*2D_ Full_ End* and *2D_ Full_ Plat*), and the larger plateaus are obtained for *Capus_ C2_ 0* and *Capus_ 2.5*. This last observation is explained since the R function *DDSE* algorithm uses a robust processed linear regression leading to plateaus with large sizes. The robustness aspect is lost on our proposed algorithms.
According to Figures from 5.12 to 5.15, the slope between our proposed minimal penalties and the least-squares values are 1 for large models when the model collection is on the entire dataset without resampling (*path*), meaning that hyperparameters are calibrated to satisfy the expected theoretical affine behavior. However, this is not the case when resampling is considered to generate the model collection since the black lines do not overlap with the blue curves anymore. For these types of model collection, the numbers of models used to calibrate hyperparameters $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ are higher and plateaus are larger than those when the collection is only on the entire dataset (Figures from 5.16 to and 5.21).

To conclude this part, the expected affine behavior is still visible with our proposed minimal penalty functions although it is less obvious to detect by our algorithms than the existing ones.

5.6.3 Predictive risk

This subsection is devoted to analyze performances of our proposed algorithms in comparison to the existing ones: *Capus_ C2_ 0*, *Capus_ 2.5* and *LinSelect*. Comparisons are made through

the study of predictive performances. The mean squared errors (MSE) (5.18) are computed in practice with $m = \hat{m}_{(\widehat{C_1}(\sigma^2),\widehat{C_2}(\sigma^2))}$ to evaluate the predictive risk.

In Figures 5.3 and 5.4, the MSE values obtained respectively from the cluster and the scale-free-max designs are summarized. As the MSE is normalized, a value smaller than 1 leads to prediction performances: the selected variables predict Y better than an empty set of variables.

First of all, the *grid* approaches have MSE values always smaller than all the other methods, especially the existing ones: LinSelect, *Capus_C2_0* and *Capus_2.5*. It proves that there always exists a couple of constants in $[0, 10]^2$ leading to better prediction performances than when there is only one hyperparameter to estimate.

When the model collection is generated on the entire dataset without resampling, *Capus_2.5* is unsurprisingly better than *Capus_C2_0* but also better than our proposed methods. Concerning the scale-free-max design, all medians are significantly below 1, meaning that the methods have all prediction performances. Medians are equals to 0.46, 0.53, 0.43, 0.45 for respectively *2D_Stop_End*, *2D_Stop_Plat*, *2D_Full_End* and *2D_Full_Plat*, higher than the *Capus_2.5* one (0.35), while the *grid* one equals 0.32. All our proposed methods are better than *LinSelect* one (0.54). Concerning the cluster design, medians are equals to 0.99, 1.07, 1.11 and 0.99 for respectively *2D_Stop_End*, *2D_Stop_Plat*, *2D_Full_End* and *2D_Full_Plat*, all larger than the *LinSelect* and the *Capus_2.5* ones (0.97), while the median of *grid* equals 0.87. However, in average, *2D_Full_End* and *2D_Full_Plat* are better than *Capus_2.5* with mean values respectively equals 1.27, 1.09 and 1.58, while the *grid* one equals 0.87. Concerning *2D_Stop_End* and *2D_Stop_Plat*, there are some large values: until 862 and 111 with an average equals 46 and 5 for respectively scale-free-max and cluster designs. Fortunately, the median values close to 1 show that these outliers only concern a few iterations.

When resampling is added for the model collection, excepted with *path_Bolasso_LS_each_resample* on the scale-free-max design where our proposed methods lead to values higher than the *Capus_2.5* one, both *Capus_C2_0* and *Capus_2.5* are the worst methods. Their MSE medians even exceed 1 in the *cluster* design. Hence, when resampling is processed, these existing methods calibrating only one constant are not predictive. On scale-free-max design, our proposed methods are always smaller than 1 (median and mean): they are predictive in this setting.
Concerning the *LinSelect*, sometimes our proposed methods are better, sometimes worse. More precisely, with *path_Bolasso_LS_each_resample*, all of our methods values are smaller than the *LinSelect* one, which is the worst method with still relatively low value: 0.61. Hence, all methods are predictive. Among our proposed strategies, the best one is *2D_Stop_End* with a median equal to 0.44. As for the cluster design, the best one is *2D_Full_End* with a median equal to 0.99, as the *LinSelect* one, and with the smallest variability. However, on average, the MSE mean exceeds 1. The other new proposed strategies lead to medians and means larger than 1.
With *path_Bolasso_LS_fulldata*, *2D_Stop_End* and *2D_Stop_Plat* lead to the best prediction performances (0.52 and 0.46 respectively) on the scale-free-max design, while the *LinSelect* median is 0.60. On the cluster setting, among our proposed strategies, *2D_Full_End* corresponds to the smallest median (0.99) and the smallest variability. Our other ones lead to medians and means larger than 1. All four are worse than *LinSelect* (median= 0.97).

With *path_Bolasso_all_models*, *2D_Full_Plat* is the best method among the new proposed strategies on the scale-free-max design, but all lead to median values smaller than 0.5 and are better than *LinSelect*. As for the cluster setting, *2D_Stop_End* and *2D_Stop_Plat* strategies are predictive: their MSE medians are smaller than 1. Moreover, *2D_Stop_End* median value is smaller than the *LinSelect* one.

Finally, in global comparison, *2D_Stop_End* with *path_Bolasso_all_models* leads to the best prediction performance for both scale-free-max and cluster designs. Indeed, its median and mean are smaller than 1 and smaller than the *LinSelect*, *Capus_C2_0* and *Capus_2.5* ones. Moreover, its variability is small (standard deviation values 0.09 and 0.12 for respectively scale-free-max and cluster).

To conclude, although *Capus_2.5* remains the best method for prediction without resampling process during the model collection generation, our proposed strategies improve prediction performances in comparison to the existing *LinSelect*, *Capus_C2_0* and *Capus_2.5* penalties when a resampling procedure is added. For the resampling process, except for *2D_Stop_End* with *path_Bolasso_all_models*, all the other combinations of model selection strategy and model collection generation lead to medians larger than the *LinSelect* one and often larger than 1, in the cluster design. Conversely, our proposed methods are always better than the *LinSelect* one and always smaller than 0.5 on the scale-free-max design. Hence, one of our proposed strategies improves the prediction performance compared to all existing methods in cluster and scale-free-max designs: it is the combination of *path_Bolasso_all_models* and *2D_Stop_End*.

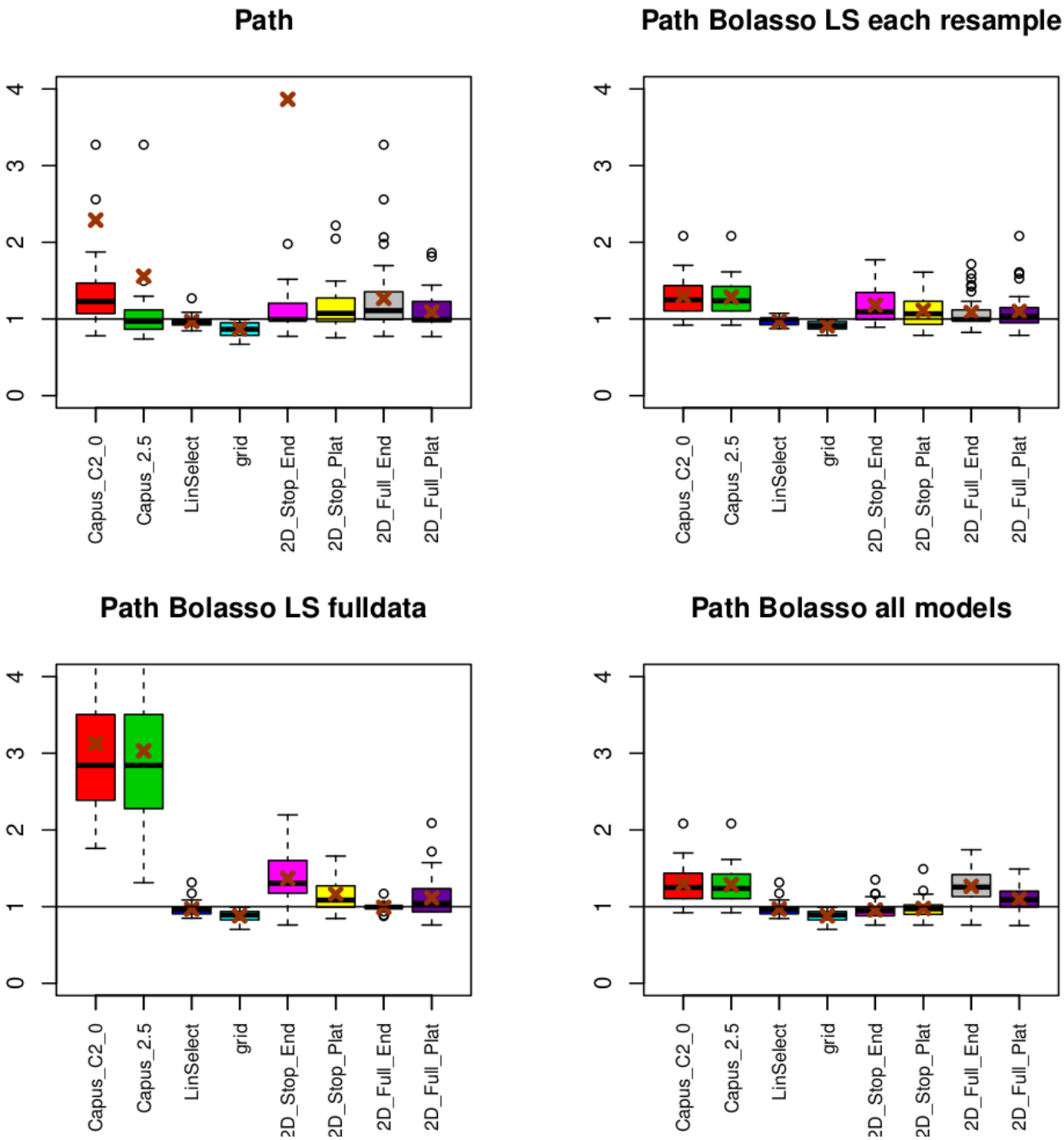


Figure 5.3: Boxplots of the MSE values obtained with the cluster setting. The brown crosses indicate the average values per penalization function (if the value is not displayed, it is out of the visible frame). A MSE value lower than 1 means that the method has a prediction performance: the selected variables predict Y better than an empty set of variables.

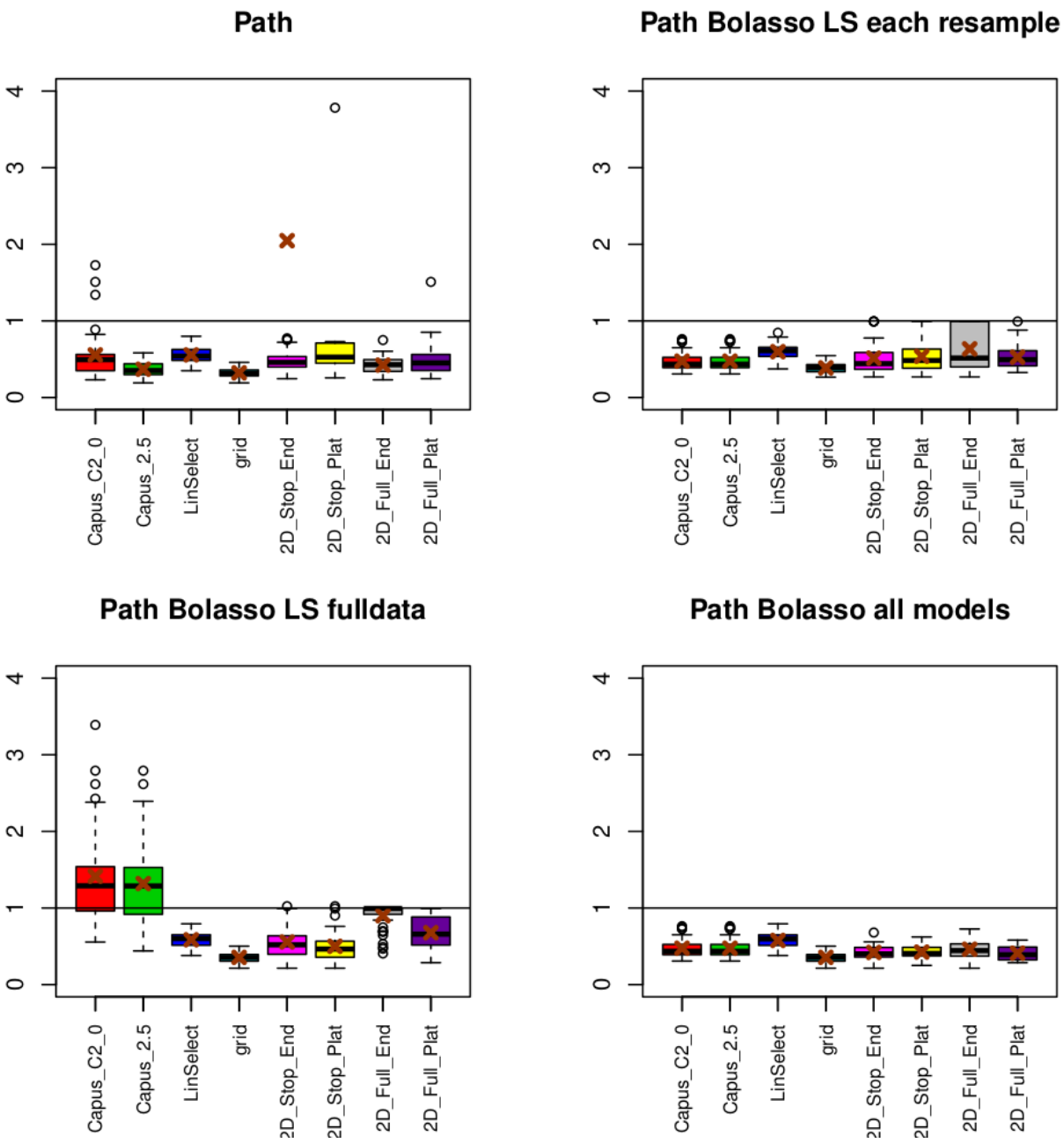


Figure 5.4: Boxplots of the MSE values obtained with the scale-free-max setting. The brown crosses indicate the average values per penalization function (if the value is not displayed, it is out of the visible frame). A MSE value lower than 1 means that the method has a prediction performance: the selected variables predict Y better than an empty set of variables.

5.7 Conclusions

Variable selection procedures by penalized criteria minimization are commonly used to correctly predict the response variable Y in the Gaussian linear regression model with a high-dimensional context, unknown variance, and dependencies between variables. With the additional sparsity assumption, we focus on the model selection method. In this chapter, the model collection is generated from a regularization path obtained by the regularized least-squares minimization. That leads to a random and data-dependent model collection. The model selection procedure involves selecting the best model among the model collection. We focus on optimal penalty functions leading to non-asymptotic guarantees on the predictive risk control. On the one hand, the LinSelect penalty guarantees a predictive risk control under the assumption of the randomness of the model collection but is a completely deterministic function. On the other hand, the data-driven penalties via the slope heuristics are data-dependent but theoretical properties are known only for a deterministic model collection.

The optimal data-driven penalty function considered in this work depends on two unknown hyperparameters $\Big(C_1(\sigma^2), C_2(\sigma^2)\Big)$. We propose new algorithms based on the slope heuristics method to calibrate them in an automatic way. We are inspired by the existing slope estimation algorithm that we improve in different ways. Firstly, we generalize the algorithm of the R function *DDSE* to the calibration of two hyperparameters simultaneously. These parameters are estimated directly on the available dataset by exploiting the expected theoretical affine behavior between the least-squares and the minimal penalty values. Instead of using a robust linear regression to estimate the slope coefficient (as in the R function *DDSE*), a constrained minimization procedure calibrates the coefficients of the best affine trend in two dimensions. Secondly, while the slope heuristics method is theoretically based on an exhaustive exploration of all possible models, we propose to consider the randomness of the model collection in our proposed algorithm. Our approach consists in adding a resampling procedure in the model collection generation. Hence, our proposed data-driven variable selection procedure is data-dependent in both model collection and model selection.

We propose several types of model collection and four data-driven penalty calibrations based on the slope heuristics principle. A simulation study based on datasets generated from Gaussian graphical models evaluates combinations of model collections and data-driven penalty calibrations. The expected affine behavior still exists for large models and our new procedures seem to differ from the existing slope heuristics methods. The simulation study shows that whatever the setting and model collection types, there always exists a couple of constants (C_1, C_2) leading to better prediction performances than with the LinSelect penalty and with the existing slope heuristics. Hence, a minimal penalty function with two unknown hyperparameters is preferred to a minimal penalty function with only one unknown hyperparameter. Adding a resampling procedure allows our data-driven penalty calibrations to improve prediction performances from no resampling model collection generation. Moreover, our proposed algorithms lead to smaller MSE values than the existing slope heuristics. The combination of the model collection type *path_Bolasso_all_models* and the data-driven penalty calibration *2D_Stop_End* improves

the prediction performance compared to all existing methods in both cluster and scale-free-max designs. Hence, to stop the model collection exploration at $D_m = \frac{p}{2}$ seems to be the best strategy.

5.8 Perspectives

First, our proposed methods should be tested on other data sets to complete our simulation study. For instance, the independent and the scale-free-min dependency structures in the Gaussian graphical model could be considered for comparison. According to Chapter 3, it is expected that results from the independent setting would be similar to those from the scale-free-max one, whereas results from the scale-free-min setting would be similar to those from the cluster one. Moreover, in future work, more different sparsity degrees could be added in the comparison, as well as other values for n and p or the value of α in the Elastic-Net criterion 5.17. In particular, if the support size of β^* is large enough, the expected theoretical affine behavior will only concern a few models at the end of the collection. Calibrating two hyperparameters would be a complex problem with few triplets and existing slope heuristics methods should be preferred. So, the procedure could be adapted according to the support size of β^*.

Only the MSE metric is tested in the comparison. However, other performances would be interesting to evaluate. As in Chapter 3, we could include the recall, the specificity and the FDR metrics in the comparison study. In particular, according to Subsection 5.6.2.3, estimated hyperparameters from our algorithms are higher than from state of the art. Thus, our penalty functions take higher values leading to more conservative procedures than the existing ones. So, the FDR metric will probably be smaller for our proposed methods than the existing ones.

One of our proposed strategies improves the prediction performance compared to all existing methods: the combination of *path_Bolasso_all_models* and *2D_Stop_End*. Its medians (resp. mean) are 0.94 and 0.40 (resp. 0.96 and 0.42) for respectively the cluster and scale-free-max design. Those of the *grid* approach are 0.89 and 0.35 (resp. 0.87 and 0.36) respectively the cluster and scale-free-max design. Hence, our best combination is still far from the naive approach which can be seen as the best we can do in our context and with the penalty (5.9). One of the main directions to improve this method would be to add robustness, which was lost by constrained minimization. Indeed, noise at the end of the model collection is not removed and as soon as a model does not belong to the affine behavior is considered to estimate the hyperparameters $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$, the slope changes and so the corresponding selected model changes too: plateaus are more numerous and of shorter length. Consequently, our strategies are very sensitive to the "*pct*" input parameter: a slight variation of its value leads to a completely different selected model. Adding robustness can stabilize the model selection procedure and be processed by improving the *solve.QP* R function of the R package *quadprog* through, for instance, a weighted least-squares minimization as in the R function *rlm*; or by using a completely different R function. Lastly, improving our algorithms has to focus on the computational time which is currently important. For instance, the impact of resampling could be investigated to get the smallest number of required resamples.

Moreover, it is possible to find a new strategy different from ours to calibrate the hyperparameters $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ automatically when resampling to construct the model collection is taken into account. A numerical process or a deterministic procedure could be put in place to distinguish an affine relationship from a non-affine relationship. Firstly, by analyzing the 3D-regularization path curves from the entire dataset of the resamples (see Figure 5.5), we observed that in the over-fitting area, regularization path curves from resamples are far below the one from the entire dataset. In contrast, in the small models area, the regularization path curves from resamples are close to and below or above the one from the whole dataset. Hence, this interesting phenomenon could be a criterion for identifying the over-fitting area, where the affine behavior will be satisfied. Secondly, as we generalize the slope coefficient estimation to

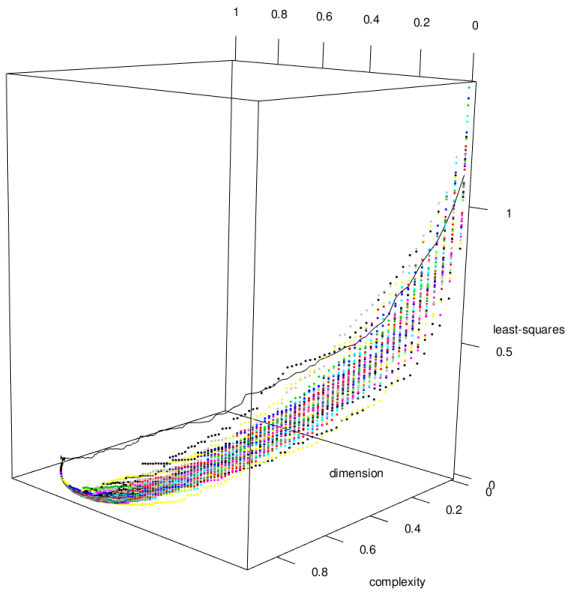


Figure 5.5: Examples of the $-\gamma_n\left(\hat{\beta}_m\right)_{m\in\mathcal{M}(\mathcal{D})}$ values as a function of the $\left(\frac{D_m}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ and $\left(\frac{\log(\binom{p}{D_m})}{n}\right)_{m\in\mathcal{M}(\mathcal{D})}$ ones. The black solid line stands for the model collection $\mathcal{M}(\mathcal{D})$ on the entire dataset. Each dotted line stands for a model collection $\mathcal{M}_r(\mathcal{D}^r)$ on a resample $\mathcal{D}^r$.

the coefficients estimations of the best affine trend in two dimensions, considering the minimal penalty 2.24 with three hyperparameters to calibrate could be an extension of this work.
Thirdly, for several iterations, $\widehat{C_2}(\sigma^2) = 0$ is obtained, meaning that the logarithm term does not appear anymore in the minimal penalty. It is undesirable since, in our framework, the logarithm term is theoretically essential. Hence, imposing a positive value for $\widehat{C_2}(\sigma^2) = 0$ could improve the model selection procedure.
Lastly, we could compare our methods with the slope heuristics strategy of [Meynet and Maugis-Rabusseau, 2012]. If this method is competitive, one prospect would be to add our constrained minimization principle in the existing slope heuristics calibrating only one hyperparameter.

Finally, Subsection 5.6.2.3 emphasizes that the ratio $\frac{C_1(\sigma^2)}{C_2(\sigma^2)}$ is different from 2.5 and is not constant at all. From a theoretical point of view, we could ask ourselves if there is not an explicit data-dependent function for it. In this case, the existing slope heuristics calibrating only one hyperparameter could be used by introducing the explicit function of the ratio in the algorithm.

5.9 Appendix: Proof of the existence and the uniqueness of the solution in the minimization under constraints problem (5.16).

Proof. With the notation of Proposition 5.1, the minimization problem (5.14) can be rewritten by the minimization of the following function G with respect to (C_1, C_2, C_3):

$$G(\theta) = n||y - (C_1 x + C_2 z + C_3 \mathbb{1}_d)||_{2,n}^2$$

Moreover, for all (C_1, C_2, C_3):

$$\begin{aligned} G(\theta) &= n||y - W\theta||_{2,n}^2 \\ &= y^T y - y^T(W\theta) - (W\theta)^T y + (W\theta)^T(W\theta) \\ &= y^T y - 2y^T W\theta + \theta^T W^T W\theta \end{aligned}$$

Hence, minimizing G with respect to θ is equivalent to minimizing J defined in (5.15). So, solving (5.14) is equivalent to solve the problem of minimization under constraints (5.16).

To prove the existence and the uniqueness of the solution of the (5.16) problem, we use the following theorem:

Theorem 5.1 ([Beck et al., 2005]). *Let us consider $X_{ad} \subset \mathbb{R}^d$ a subset of $\mathbb{R}^d$ for $d \geq 1$, and $J : X_{ad} \mapsto \mathbb{R}$. If*

(i) X_{ad} is a closed subset of $\mathbb{R}^d$

(ii) J is inf-compact on X_{ad}

then the minimization problem of J on X_{ad} has a solution (existence).
If in addition X_{ad} is convex and J strictly convex, then the solution is unique.

Existence of a solution of (5.16).
(i) the space $X_{ad} = \left\{\theta, A^T\theta \geq b_0\right\}$ is a closed subset of $\mathbb{R}^3$ since X_{ad} is the inverse image of the closed subset $\left\{(0,0,0)^T\right\}$ under a linear map.
(ii) To prove that J is inf-compact on X_{ad}, we use the following proposition:

Proposition 5.2. *In finite dimension, if J is coercive and lower semicontinuous on X_{ad}, then J is inf-compact on X_{ad}.*

Here, J is continue on X_{ad}, so is semicontinuous on X_{ad}. J is coercive since:

$$\begin{aligned} J(\theta) &= \frac{n}{2}||W\theta||_{2,n}^2 - \langle u, \theta\rangle \\ &\underset{(**)}{\geq} \frac{n}{2}\lambda_{\min}||\theta||_{2,n}^2 - |\langle u, \theta\rangle| \\ &\underset{(***)}{\geq} \frac{n}{2}\lambda_{\min}||\theta||_{2,n}^2 - n||u||_{2,n}||\theta||_{2,n} \quad \underset{\substack{\sqrt{n}||\theta||_{2,n}\longrightarrow+\infty \\ (i)}}{\longrightarrow} +\infty. \end{aligned}$$

(**) is true since W^TW is symmetric and:

$$\begin{aligned} &\theta^TW^TW\theta = 0 \iff ||W\theta||_{2,n}^2 = 0 \iff ||W\theta|| = 0 \\ &\underset{(ii)}{\iff} W\theta = 0 \iff \theta = 0. \end{aligned}$$

So, W^TW is a positive definite matrix. So, W^TW is diagonalizable on an orthogonal basis on $\mathbb{R}^3$ with strictly positive eigenvalues. Thus, $\theta^TW^TW\theta = \theta^TP^T\tilde{D}P\theta \geq \theta^TP^T\lambda_{\min}I_3P\theta = \lambda_{\min}n||P\theta||_{2,n}^2 = \lambda_{\min}n||\theta||_{2,n}^2$, with P and $\tilde{D}$ the corresponding orthogonal matrix and diagonal matrix. *(***)* comes from the Cauchy-Schwarz inequality. *(i)* comes from the fact that $\lambda_{\min} > 0$. Lastly, *(ii)* is true since W is an injective matrix.
So, $J(\theta) \underset{\sqrt{n}||\theta||_{2,n}\longrightarrow+\infty}{\longrightarrow} +\infty$ and J is coercive.
Therefore, J has a minimum on X_{ad}.

Uniqueness of a solution of (5.16).
Let $h \in \mathbb{R}^3$,

$$\begin{aligned} J(\theta + h) - J(\theta) &= 2\langle W\theta, Wh\rangle + \langle Wh, Wh\rangle - \langle u, h\rangle \\ &= 2\langle W^TW\theta - \frac{u}{2}, h\rangle + \langle Wh, Wh\rangle \end{aligned}$$

But, as $\frac{\langle Wh,Wh\rangle}{\sqrt{n}||h||_{2,n}} \leq \frac{n|||W|||_2^2||h||_{2,n}^2}{\sqrt{n}||h||_{2,n}} = \sqrt{n}|||W|||_2^2||h||_{2,n} \underset{\sqrt{n}||h||_{2,n}\longrightarrow 0}{\longrightarrow} 0$ (with the matrix norm defined in Subsection 5.1.2), then $\lim\limits_{\sqrt{n}||h||_{2,n}\longrightarrow 0} \frac{J(\theta+h)-J(\theta)-2\langle W^TW\theta-\frac{u}{2},h\rangle}{\sqrt{n}||h||_{2,n}} = \lim\limits_{\sqrt{n}||h||_{2,n}\longrightarrow 0} \frac{\langle Wh,Wh\rangle}{\sqrt{n}||h||_{2,n}} = 0$
So, $\forall h \in \mathbb{R}^3,\ dJ(\theta)(h) = \langle 2W^TW\theta - \frac{u}{2}, h\rangle$, where d designs the differential application.
Let $(h_1, h_2) \in \mathbb{R}^3$,

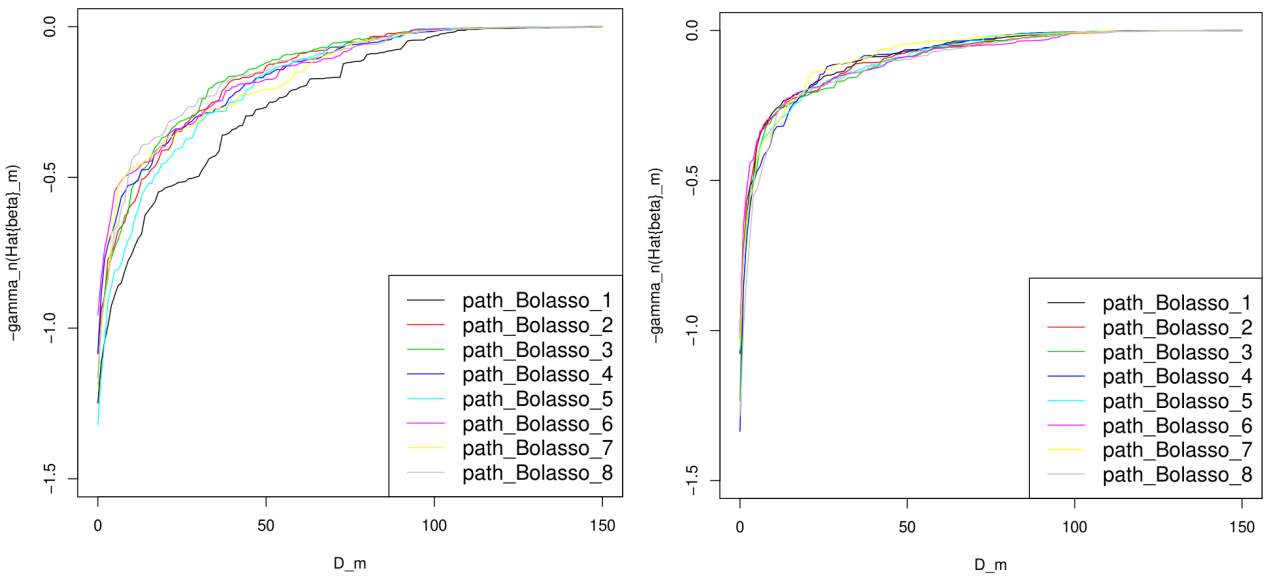
$$dJ(\theta + h_1)(h_2) - dJ(\theta)(h_2) = \langle 2W^TWh_1, h_2\rangle$$

So, $\nabla^2 J(\theta) = 2W^TW$, where ∇ designs the gradient of the second order differential application.
So, the Hessian matrix is symmetric definite positive. So, J is strictly convex.
As X_{ad} is convex, then the solution of the problem (5.16) is unique.

□

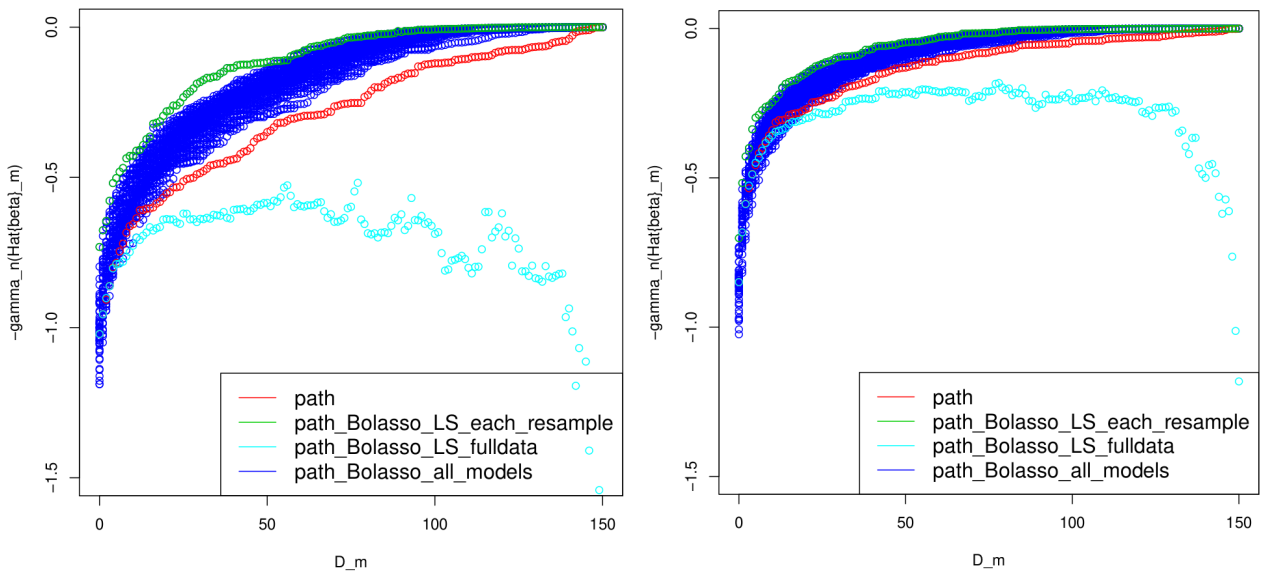
5.10 Appendix: Figures of the output characteristics of the proposed algorithms.

This subsection contains most of the figures used for analyses of the outputs of the proposed algorithms. Comments on the results can be found in Subsection 5.6.2.



cluster scale-free-max

Figure 5.6: Illustrations of the least-squares values with respect to the dimension of model collections from 8 resamples



cluster scale-free-max

Figure 5.7: Illustrations of the least-squares values with respect to the dimension of all model collections considered in this simulation study.

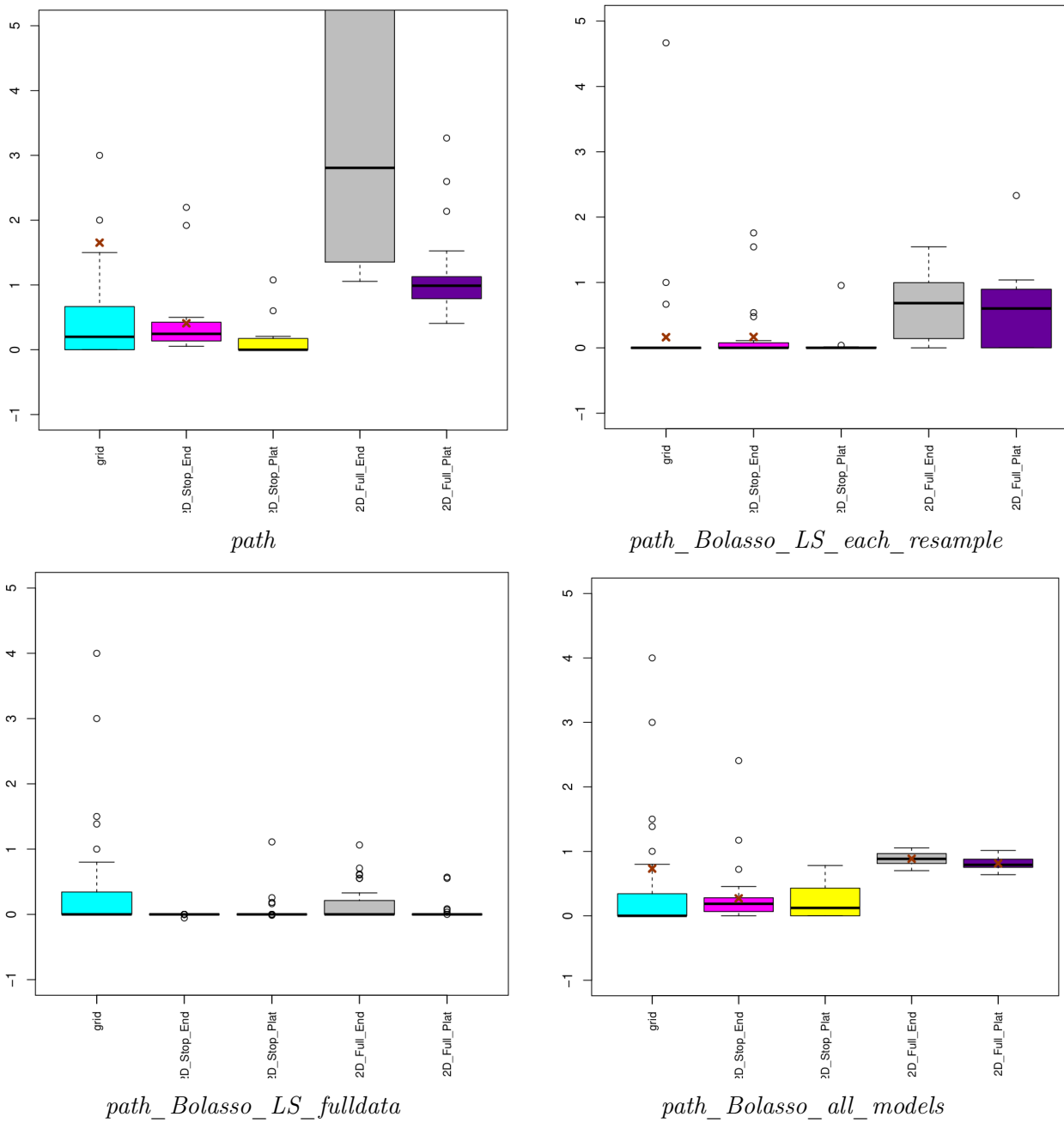


Figure 5.8: Boxplots of the $\frac{\widehat{C_1}(\sigma^2)}{\widehat{C_2}(\sigma^2)}$ ratios values obtained with the cluster setting. The brown crosses indicate the average values per penalization function.

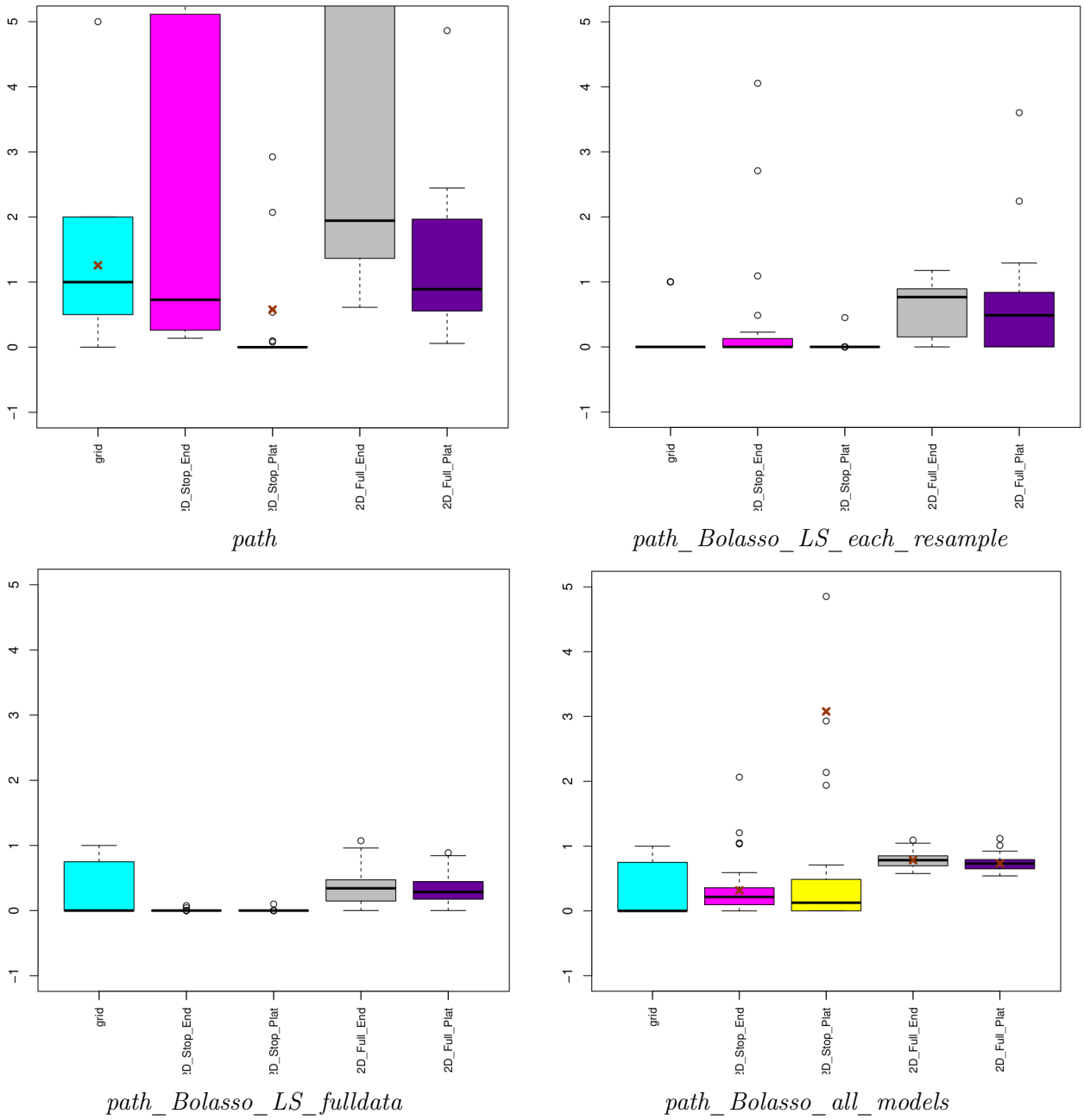


Figure 5.9: Boxplots of the $\frac{\widehat{C_1}(\sigma^2)}{\widehat{C_2}(\sigma^2)}$ ratios values obtained with the scale-free-max setting. The brown crosses indicate the average values per penalization function.

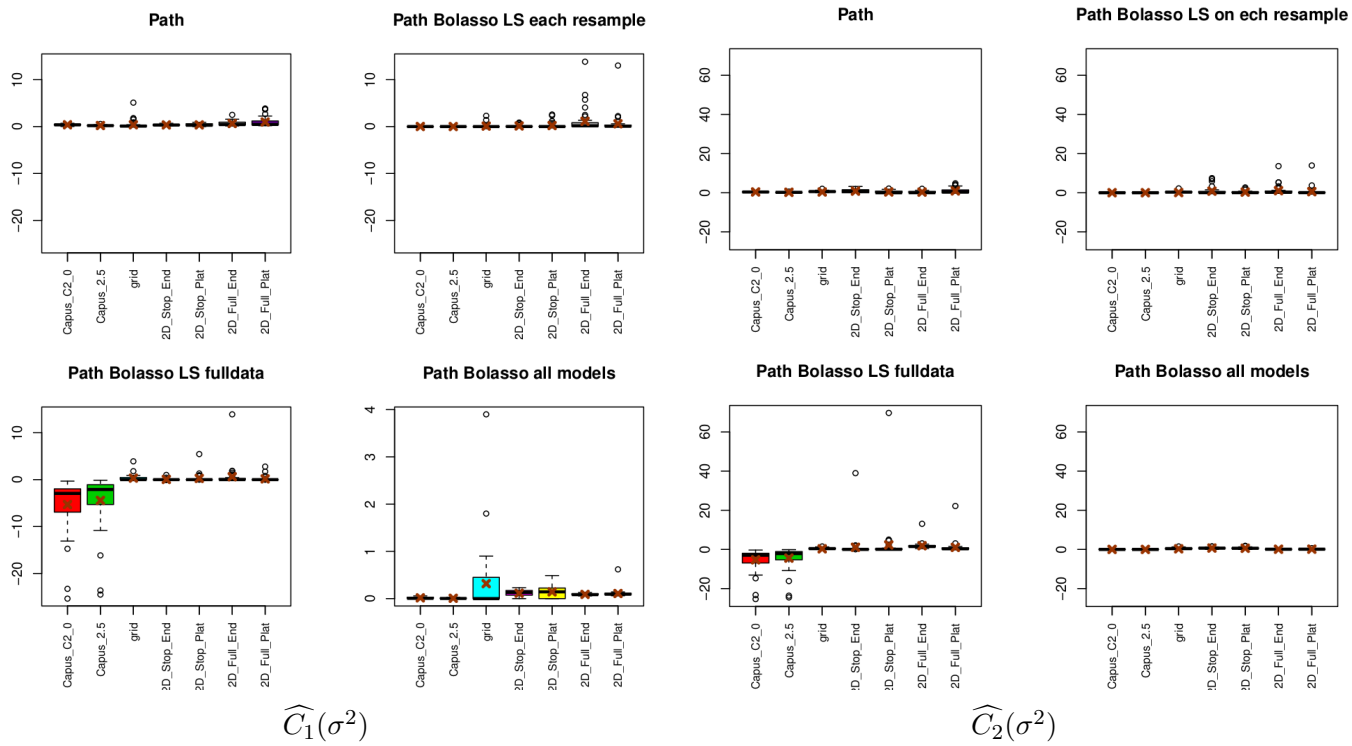


Figure 5.10: Boxplots of the $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ values obtained with the cluster setting. The brown crosses indicate the average values per penalization function (if the value is not displayed, it is out of the visible frame). Concerning the *Capus_ C2_ 0* and *Capus_ 2.5* penalties, values correspond to $\hat{\kappa}(\sigma^2)$ ones.

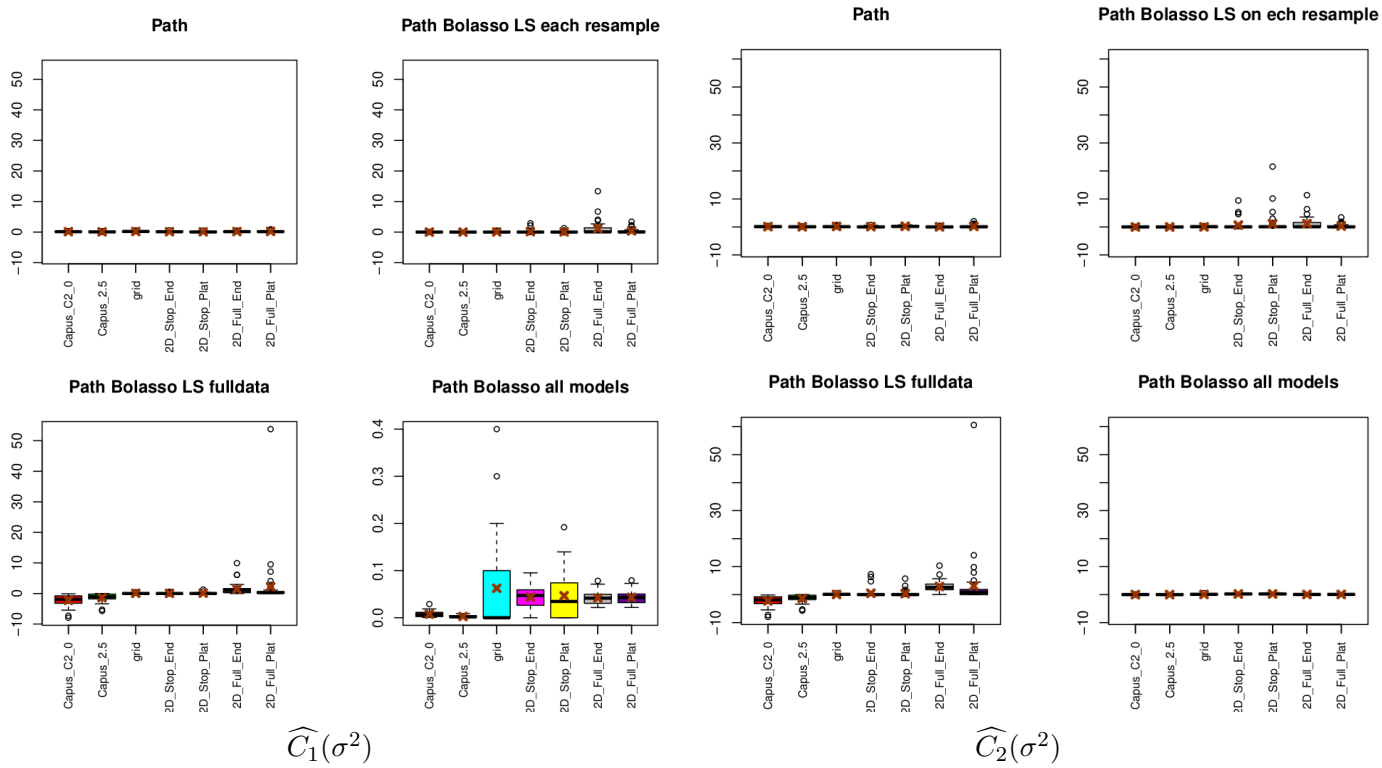


Figure 5.11: Boxplots of the $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ values obtained with the scale-free-max setting. The brown crosses indicate the average values per penalization function (if the value is not displayed, it is out of the visible frame). Concerning the *Capus_C2_0* and *Capus_2.5* penalties, values correspond to $\hat{\kappa}(\sigma^2)$ ones.

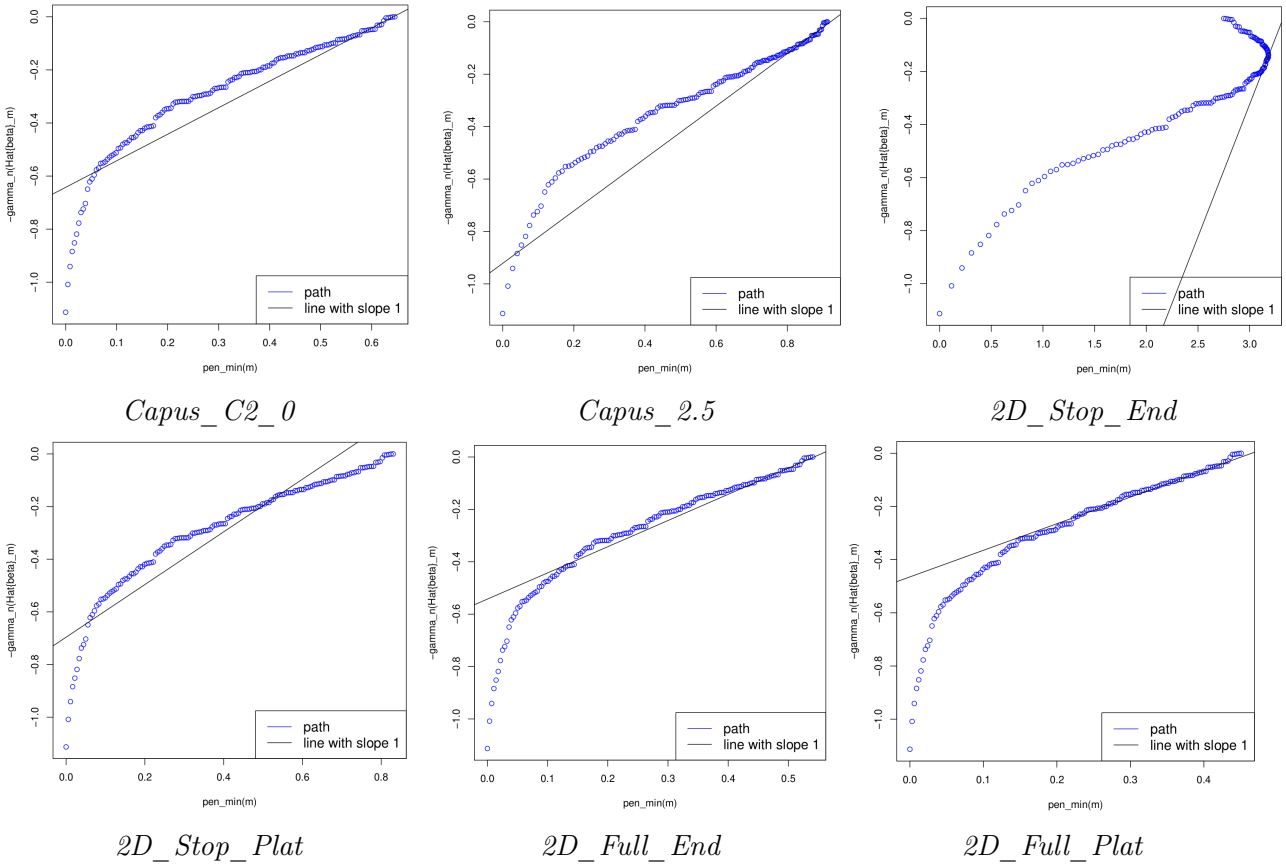


Figure 5.12: Plots of the least-squares values as a function of the minimal penalty ones. Hyperparameters $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ are calibrated according to the different methods studied when the used model collection is *path*. As an affine behavior between the $-\gamma_n\left(\hat{\beta}_m\right)$ values and the minimal penalty ones is expected for large models with a slope equals 1, the black line has to fit with the blue one for model dimension large enough. These figures concern the cluster design.

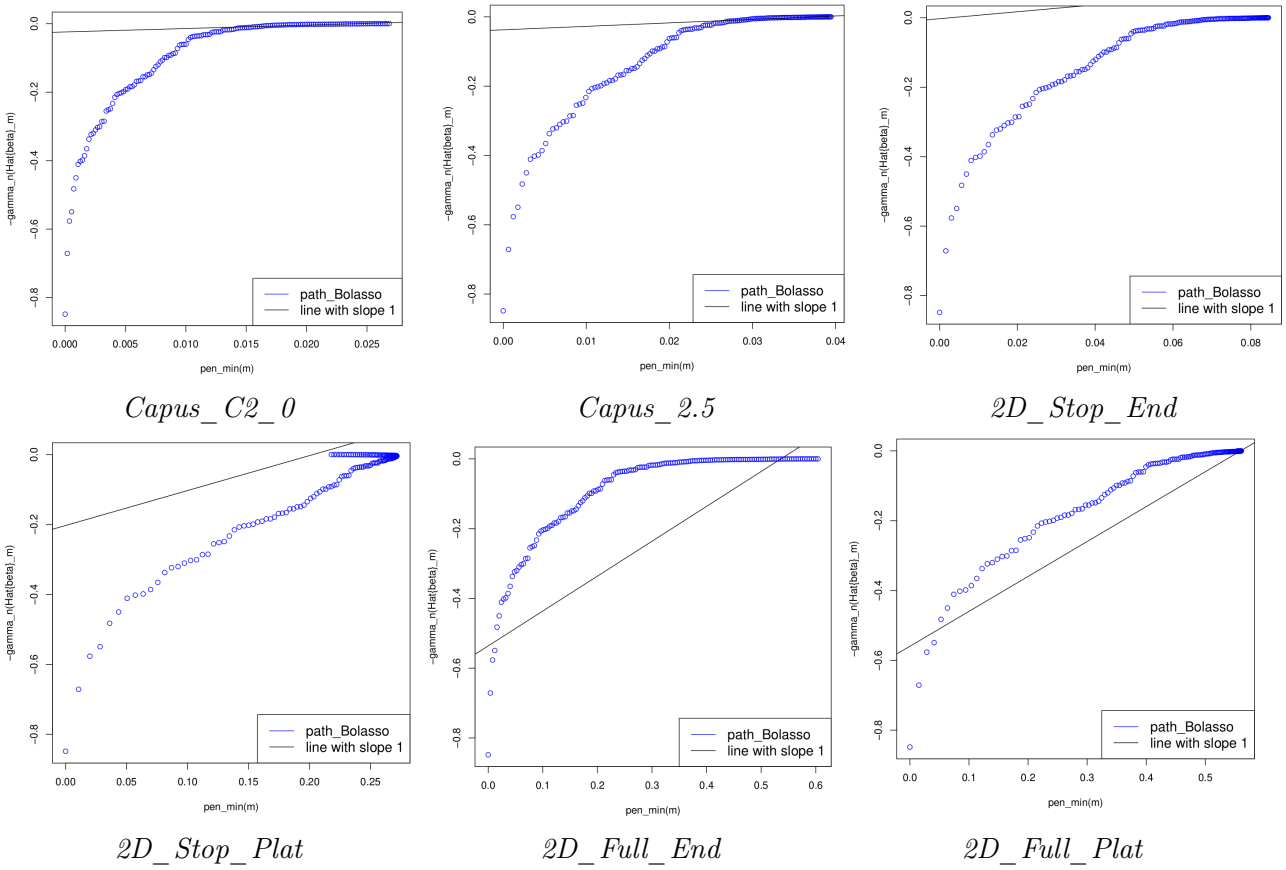
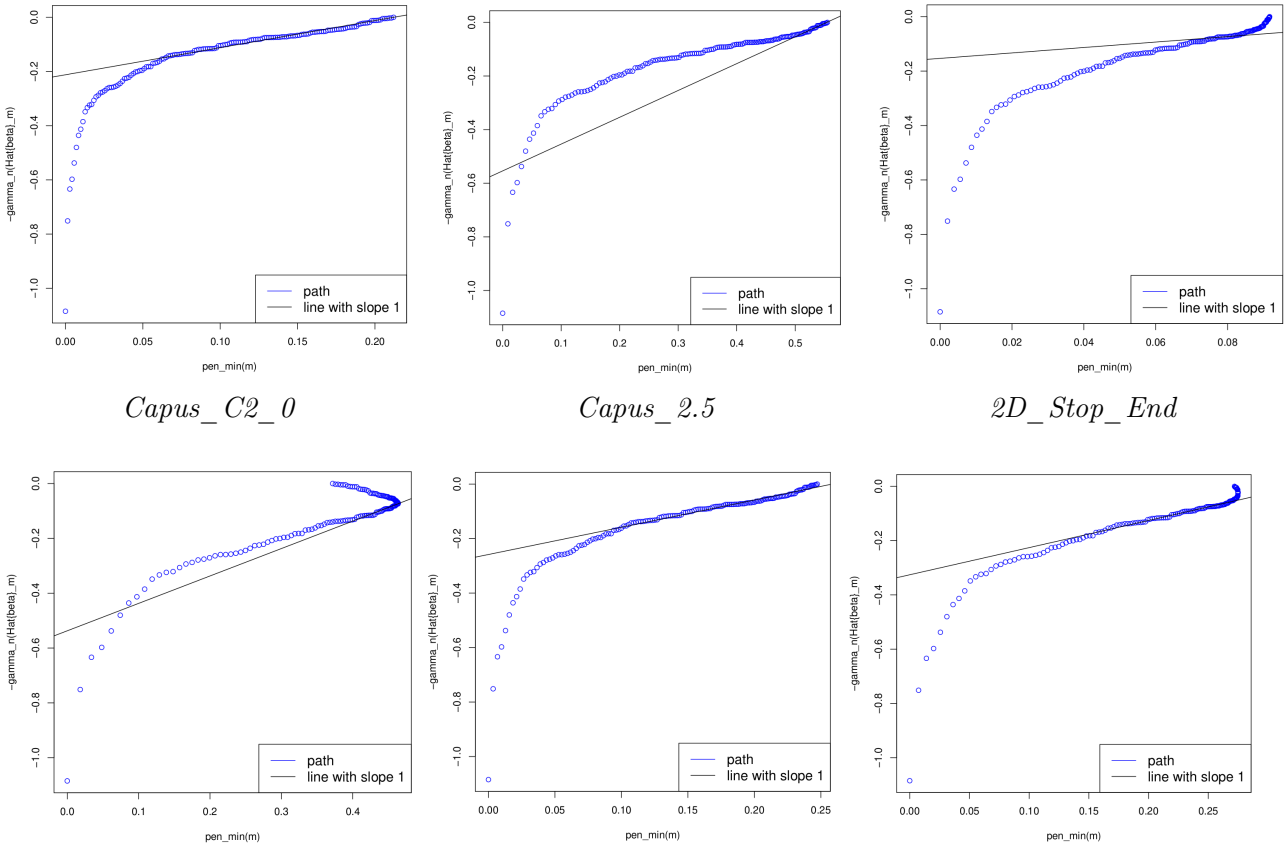


Figure 5.13: Plots of the least-squares values as a function of the minimal penalty ones. Hyperparameters $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ are calibrated according to the different methods studied when the used model collection is *path_Bolasso_LS_each_resample*. As an affine behavior between the $-\gamma_n\left(\hat{\beta}_m\right)$ values and the minimal penalty ones is expected for large models with a slope equals 1, the black line has to fit with the blue one for model dimension large enough. These figures concern the cluster design.



Capus_C2_0 *Capus_2.5* *2D_Stop_End*

2D_Stop_Plat *2D_Full_End* *2D_Full_Plat*

Figure 5.14: Plots of the least-squares values as a function of the minimal penalty ones. Hyperparameters $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ are calibrated according to the different methods studied when the used model collection is *path*. As an affine behavior between the $-\gamma_n\left(\hat{\beta}_m\right)$ values and the minimal penalty ones is expected for large models with a slope equals 1, the black line has to fit with the blue one for model dimension large enough. These figures concern the scale-free-max design.

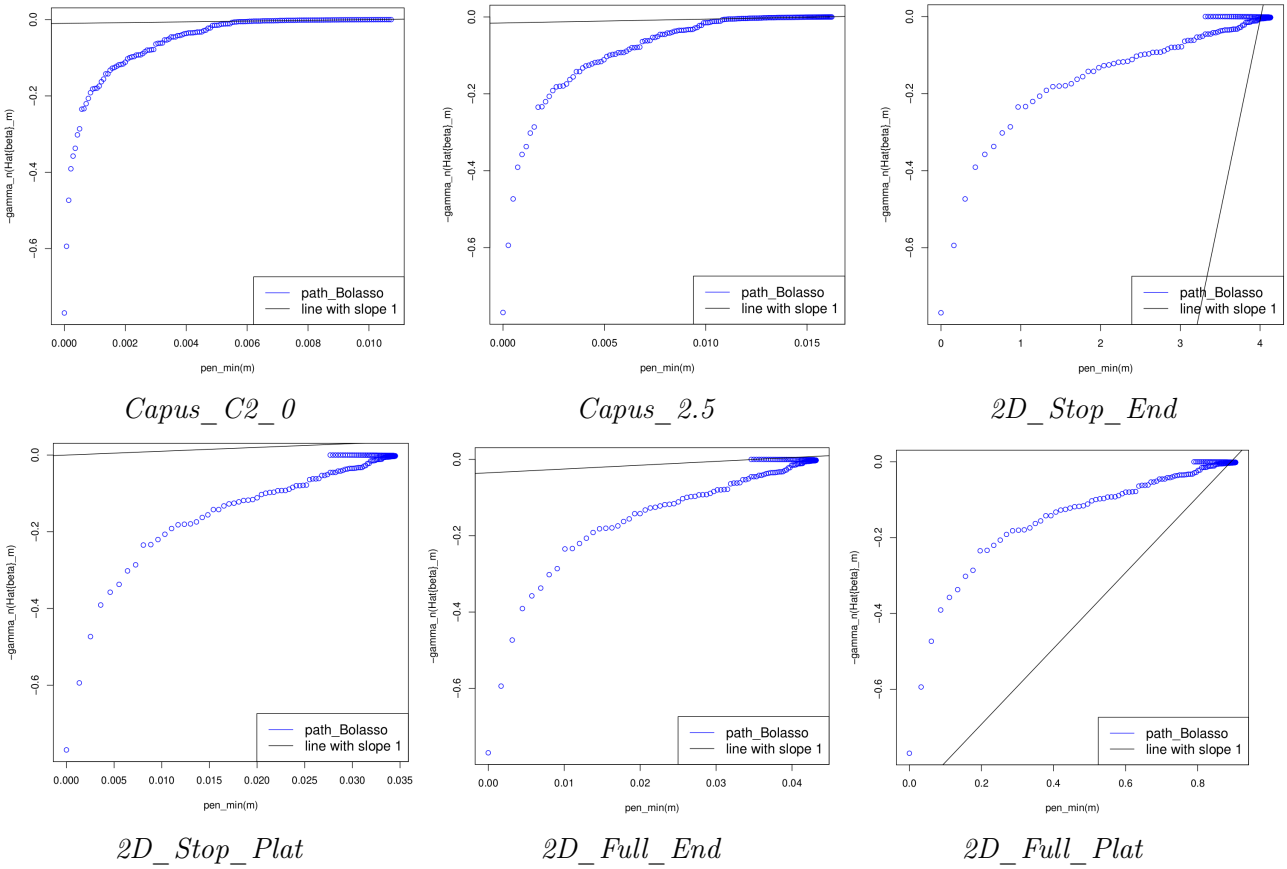


Figure 5.15: Plots of the least-squares values as a function of the minimal penalty ones. Hyperparameters $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ are calibrated according to the different methods studied when the used model collection is *path_Bolasso_LS_each_resample*. As an affine behavior between the $-\gamma_n\left(\hat{\beta}_m\right)$ values and the minimal penalty ones is expected for large models with a slope equals 1, the black line has to fit with the blue one for model dimension large enough. These figures concern the scale-free-max design.

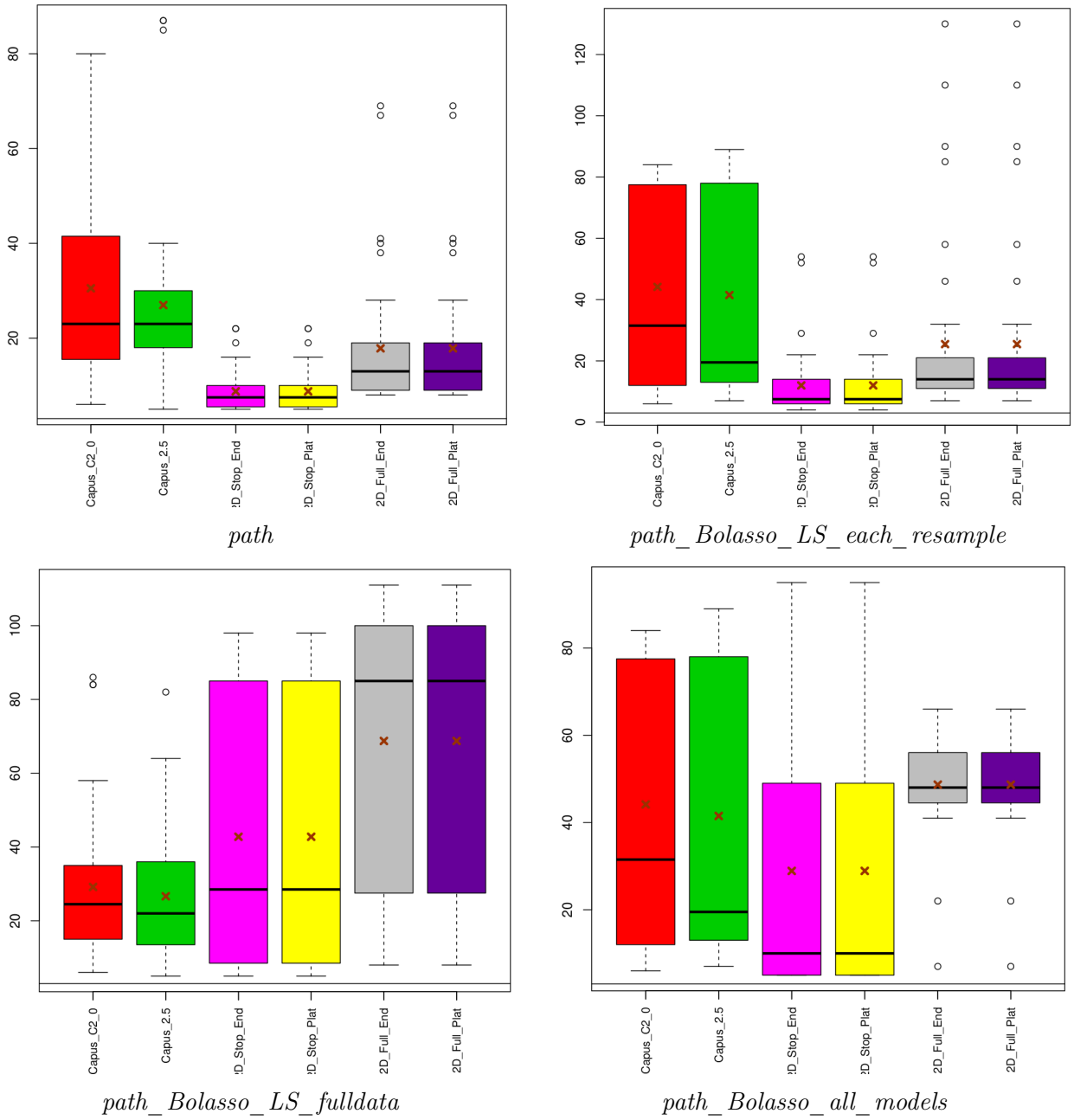


Figure 5.16: Boxplots of the number of models used to calibrate hyperparameters $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ for the cluster setting. The brown crosses indicate the average values per penalization function. The black horizontal line is the value 3 on the ordinate axis and corresponds to the minimal number of models that can be used to estimate three coefficients (the slope coefficients and the intercept).

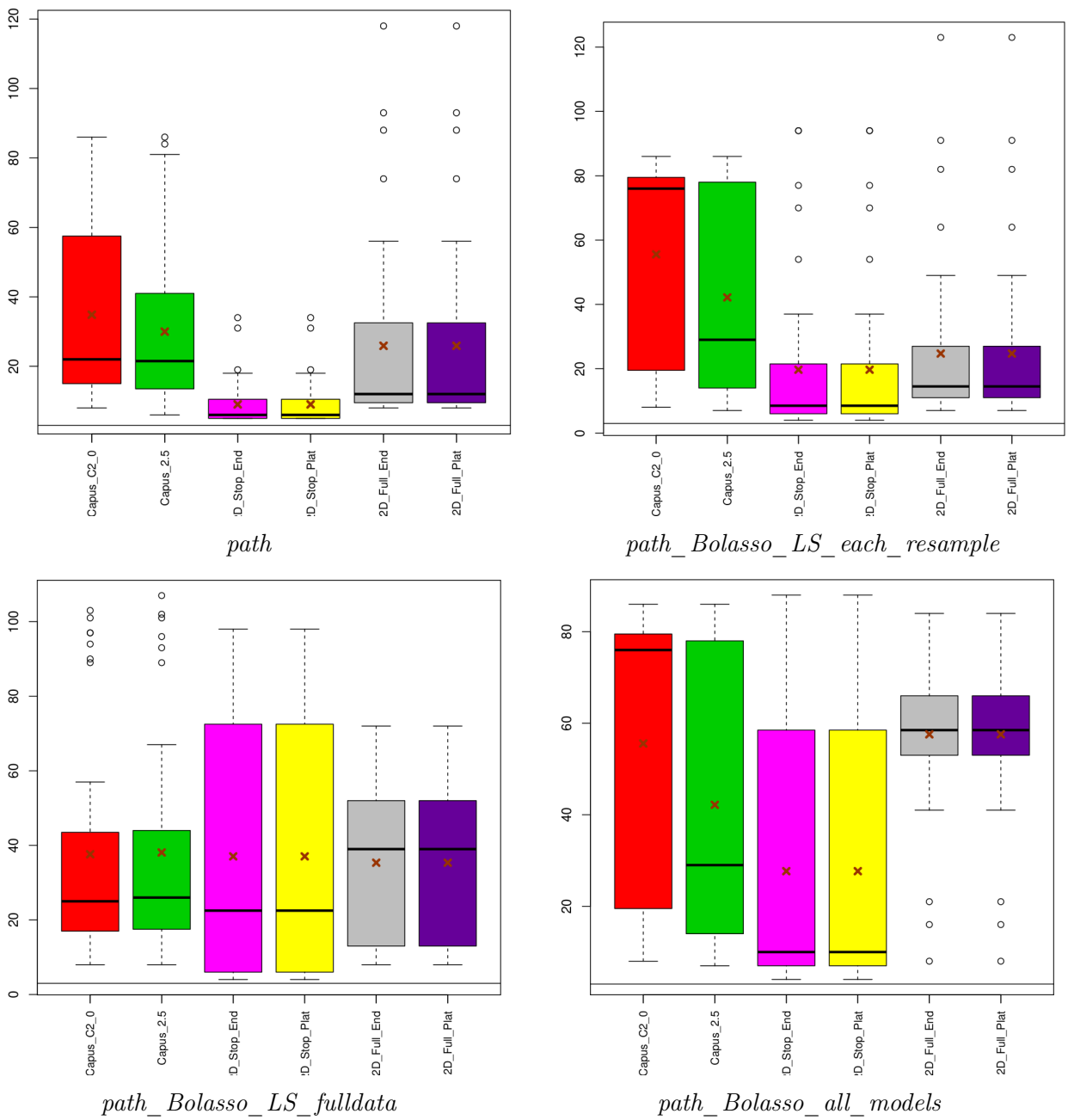
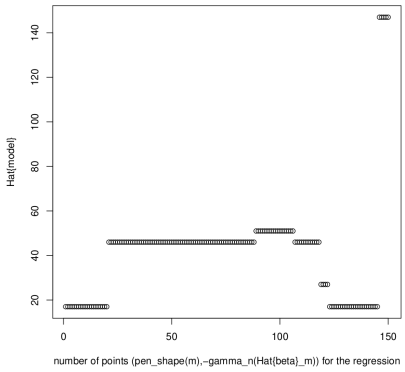
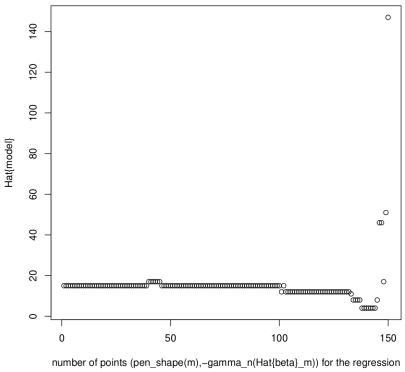


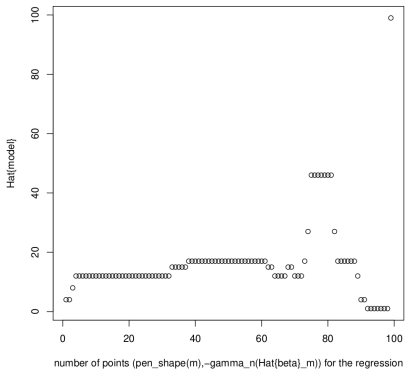
Figure 5.17: Boxplots of the number of models used to calibrate hyperparameters $\widehat{C_1}(\sigma^2)$ and $\widehat{C_2}(\sigma^2)$ for the scale-free-max setting. The brown crosses indicate the average values per penalization function. The black horizontal line is the value 3 on the ordinate axis and corresponds to the minimal number of models that can be used to estimate three coefficients (the slope coefficients and the intercept).



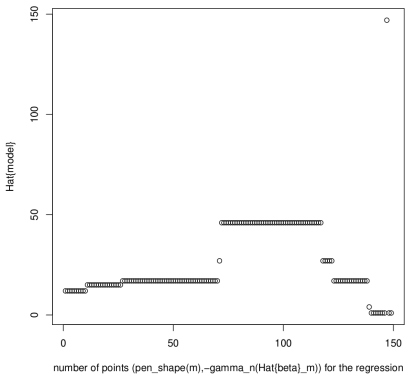
path and *Capus_ C2_ 0*



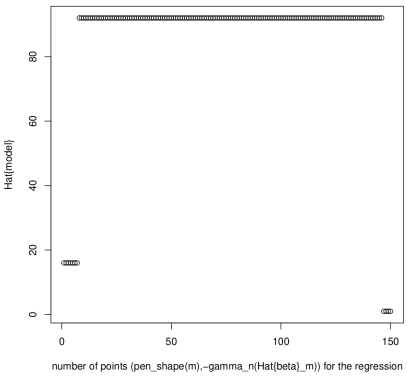
path and *Capus_ 2.5*



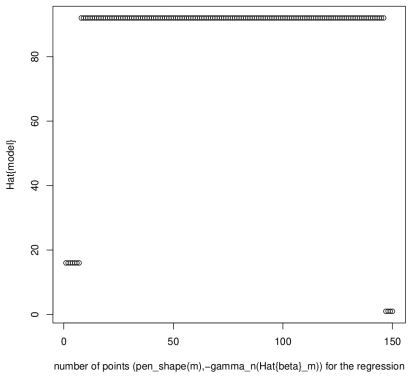
path and *2D_ Stop_ End* or *2D_ Stop_ Plat*



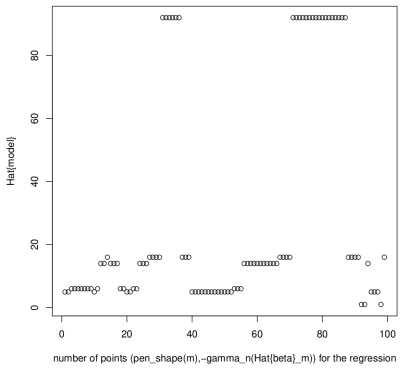
path and *2D_ Full_ End* or *2D_ Full_ Plat*



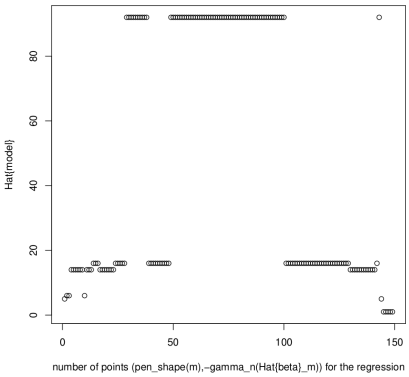
path_ Bolasso_ LS_ each_ resample and *Capus_ C2_ 0*



path_ Bolasso_ LS_ each_ resample and *Capus_ 2.5*

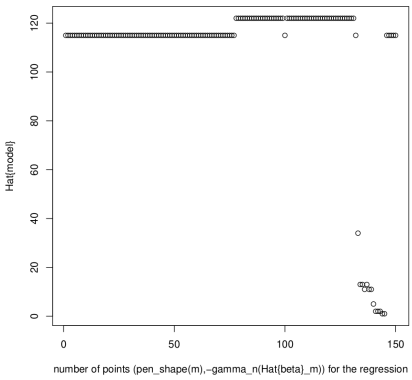


path_ Bolasso_ LS_ each_ resample and *2D_ Stop_ End* or *2D_ Stop_ Plat*

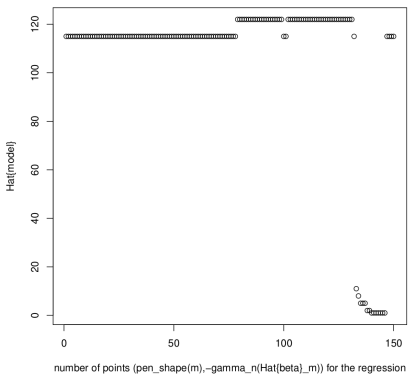


path_ Bolasso_ LS_ each_ resample and *2D_ Full_ End* or *2D_ Full_ Plat*

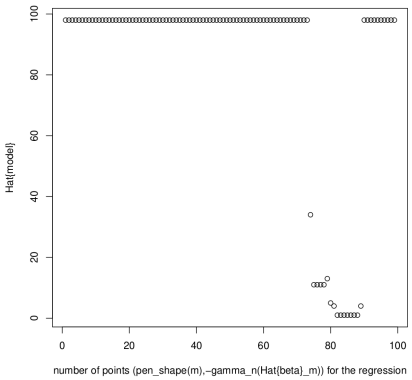
Figure 5.18: Plots of the function of the successive selected models with respect to the number of couples $\left(\mathrm{pen}_{\mathrm{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ used for the affine regression coefficients estimation. These figures concern the cluster design and model collection types *path* and *path_ Bolasso_ LS_ each_ resample*.



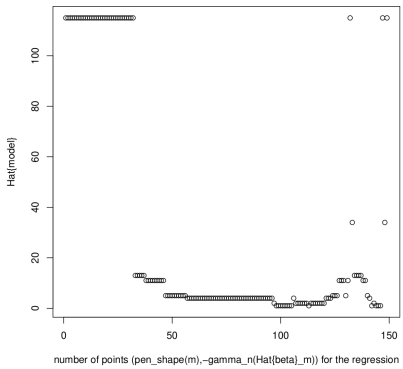
path_Bolasso_LS_fulldata and *Capus_C2_0*



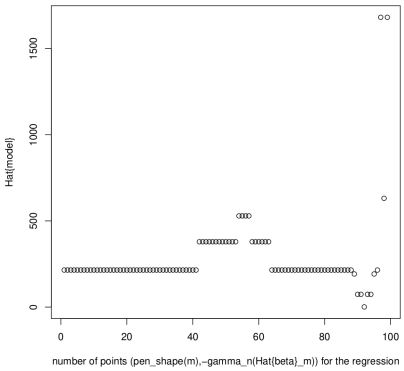
path_Bolasso_LS_fulldata and *Capus_2.5*



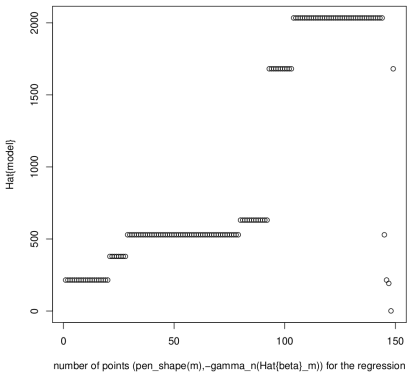
path_Bolasso_LS_fulldata and *2D_Stop_End* or *2D_Stop_Plat*



path_Bolasso_LS_fulldata and *2D_Full_End* or *2D_Full_Plat*

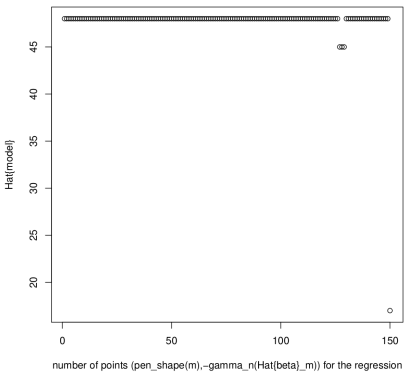


path_Bolasso_all_models and *2D_Stop_End* or *2D_Stop_Plat*

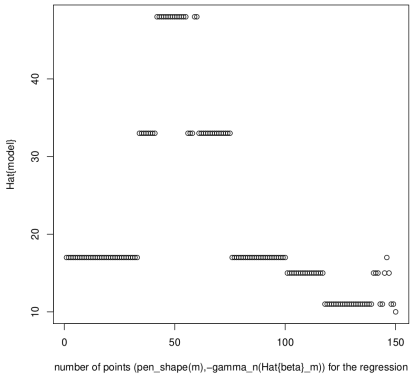


path_Bolasso_all_models and *2D_Full_End* or *2D_Full_Plat*

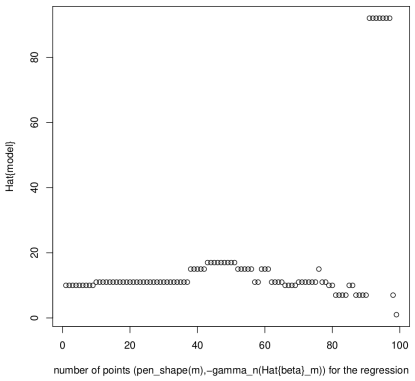
Figure 5.19: Plots of the function of the successive selected models with respect to the number of couples $\left(\mathrm{pen}_{\mathrm{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ used for the affine regression coefficients estimation. These figures concern the cluster design and model collection types *path_Bolasso_LS_fulldata* and *path_Bolasso_all_models*.



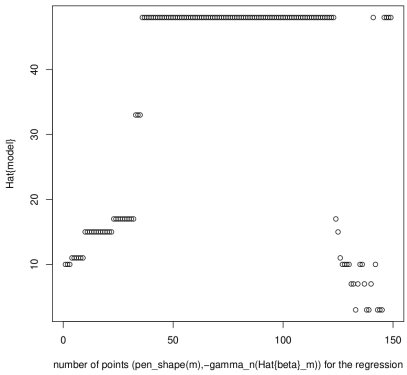
path and *Capus_ C2_ 0*



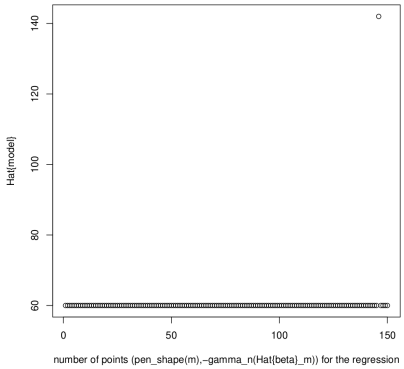
path and *Capus_ 2.5*



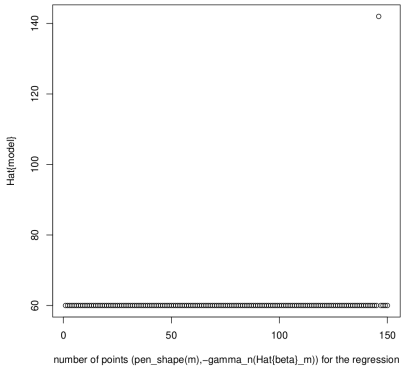
path and *2D_ Stop_ End* or *2D_ Stop_ Plat*



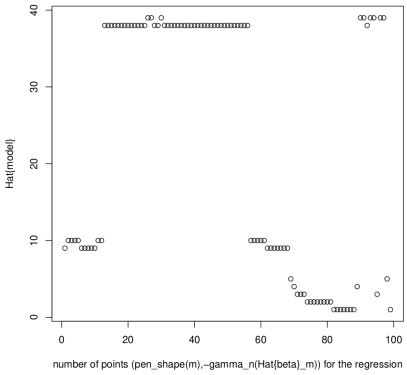
path and *2D_ Full_ End* or *2D_ Full_ Plat*



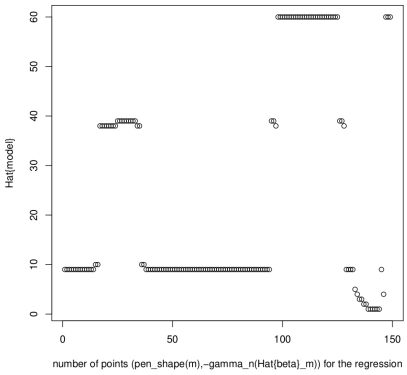
path_ Bolasso_ LS_ each_ resample and *Capus_ C2_ 0*



path_ Bolasso_ LS_ each_ resample and *Capus_ 2.5*

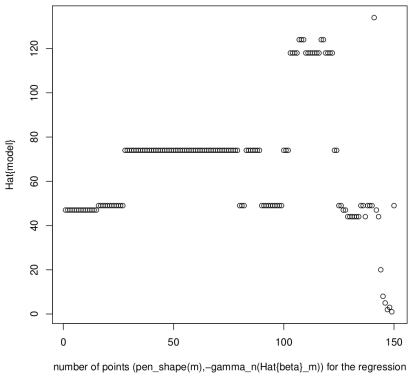


path_ Bolasso_ LS_ each_ resample and *2D_ Stop_ End* or *2D_ Stop_ Plat*

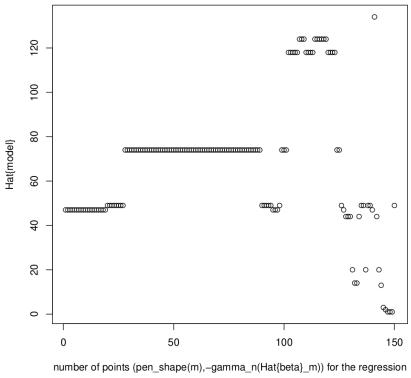


path_ Bolasso_ LS_ each_ resample and *2D_ Full_ End* or *2D_ Full_ Plat*

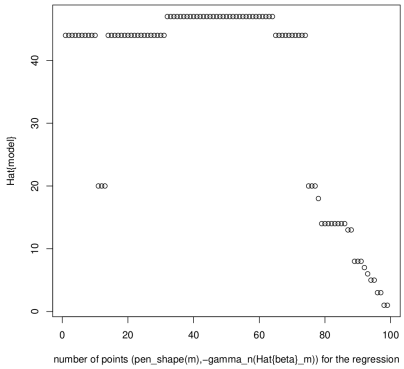
Figure 5.20: Plots of the function of the successive selected models with respect to the number of couples $\left(\mathrm{pen}_{\mathrm{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ used for the affine regression coefficients estimation. These figures concern the scale-free-max design and model collection types *path* and *path_ Bolasso_ LS_ each_ resample*.



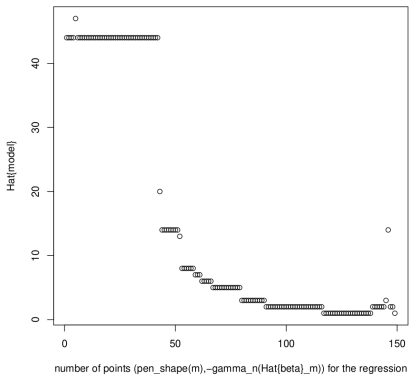
path_Bolasso_LS_fulldata and *Capus_C2_0*



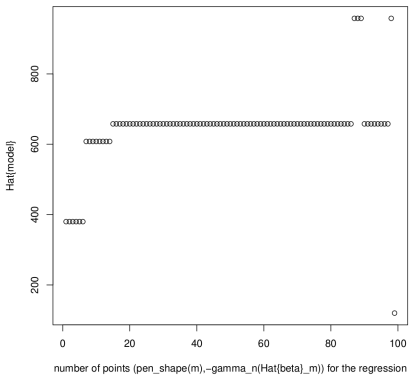
path_Bolasso_LS_fulldata and *Capus_2.5*



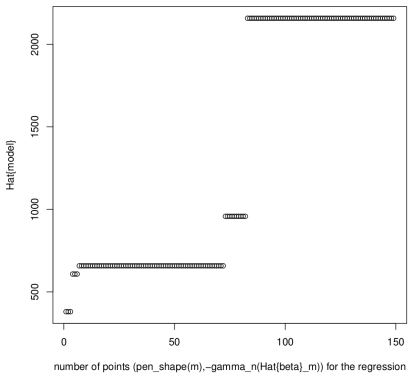
path_Bolasso_LS_fulldata and *2D_Stop_End* or *2D_Stop_Plat*



path_Bolasso_LS_fulldata and *2D_Full_End* or *2D_Full_Plat*



path_Bolasso_all_models and *2D_Stop_End* or *2D_Stop_Plat*



path_Bolasso_all_models and *2D_Full_End* or *2D_Full_Plat*

Figure 5.21: Plots of the function of the successive selected models with respect to the number of couples $\left(\mathrm{pen}_{\mathrm{shape}}(m), -\gamma_n\left(\hat{\beta}_m\right)\right)$ used for the affine regression coefficients estimation. These figures concern the scale-free-max design and model collection types *path_Bolasso_LS_fulldata* and *path_Bolasso_all_models*.

Chapter 6

Improve the regulation mechanism knowledge of genes from transcriptomic data

Abstract

The biological context motivates all the statistical problems considered in this thesis. Chapters 3, 4 and 5 consist in studying, comparing and developing statistical tools, in order to apply them on transcriptomic data to answer the biological problematic. The transition to the application on real data is complex. A pre-processing step is necessary to fit the statistical modeling as well as possible, while ensuring to lose as little information as possible from the original data. After presenting the available real data set (Section 6.2), a pre-processing and an analysis on the dataset are performed (Section 6.3). Some statistical methods (Section 6.4) are applied and the obtained selected TF sets are compared (Section 6.5). For this, we use some biological knowledge that we consider as the reference.

This work was initiated in the context of project supervision of Marion Naveau and Armand Favrot, two students of the Master "Mathématiques pour les Sciences du Vivant" of the Paris-Saclay University.

Contents

6.1 Biological question and statistical strategy

6.1.1 Recover *transcription factors* from transcriptomic data

We focus on the transcriptional activity corresponding to the first step of the gene expression. The DNA sequence is encoded into a mRNA (*messenger ribonucleic acid*) sequence, called a *transcript*. To start the *transcription* process of a gene, some proteins must be present on the *promoter region* of the gene (the DNA sequence preceding the gene). These proteins form a *protein complex*. The *transcription* process can start only when the *protein complex* is complete. These proteins themselves derive from the expression of certain specific genes that we also call, by abuse of language, *transcription factors* (TFs). Within the *protein complex*, some proteins are *activators* and promote *transcription* of the target gene, while others are *inhibitors* and prevent *transcription*. Hence, proteins from TFs regulate gene expressions and belong to the actors of the regulation mechanism.

Determining the *protein complex* formed in the *promoter region* (i.e. equivalently the associated TFs) of a target gene and the role of each protein within it helps to understand the regulation mechanism of the target gene expression. However, this is a complex problem since a target gene has several TFs; a TF can have different target genes; a TF has also several TFs (since a TF is also a target gene); and dependencies exist between genes (in particular, TFs of the same target gene are strongly dependent between themselves). So, to highlight the regulation mechanism of the target gene expression, identifying its TFs needs to take into account all the TFs simultaneously.

As proteins from TFs are actors of regulation mechanism of the target gene expression, the dataset used to recover the TFs of one target gene can be proteomic data. As example of proteomic data, the *Chromatin immunoprecipitation experiments (ChiP)* isolate a protein to identify its binding sites on the DNA, so, if this protein comes from a TF, the ChiP experiments detect its target genes. However, such protocols study TFs independently from each other and do not take into account the strong dependency between TFs. To bring a global vision to the gene expression regulation process, considering simultaneously data acquisition encompassing the entire genome is essential. Moreover, our objective is to recover the set of the TFs of one specific target gene and not to recover the target genes of one specific TF. Therefore, we turn towards a dataset produced from the *DNA microarray* technique [Lenoir and Giannella, 2006] in this thesis, giving data on the whole genome: the available *transcripts* quantities are measured on all genes simultaneously. However, there is no systematic relationship between the amount of *transcripts* and the one of proteins produced (recover the mRNA sequence from protein is impossible in general). This prevents the interpretation of the obtained set of TFs from a proteomic point of view.

Hence, the biological question we answer in this thesis is: can we recover the set of the *transcription factors* (TFs) involved in the regulation mechanism process of a given target gene through a transcriptomic data analysis?

To answer the biological question, we propose to use Gaussian linear regression modeling and apply some of the statistical methods studied in Chapter 3 as well as our proposed algorithms of 2D-slope heuristics of Chapter 5.

6.1.2 The four genes of interest

In this chapter, we are only interested in 4 target genes of *Arabidospis thaliana*: LEAFY (AT5G61850), AP1 (AT1G69120), AP3 (AT3G54340) and AG (AT4G18960). The goal is to propose a set of candidates TFs to be included in the regulation mechanism for each of the 4 target genes. For that, we use the transcriptomic dataset described in Subsection 6.2.

The four genes LEAFY, AP1, AP3 and AG are themselves TFs and have an essential role in the floral development of the plant. Moreover, they have already been extensively studied and are known biologically to interact physically with each other. Biologists expect that they regulate each other and they are TFs of the same target genes. In [Chen et al., 2018], the authors give a list of TFs for these 4 genes:

Target gene ID	Target gene name	Number of TFs	Names of TFs
AT5G61850	LEAFY	9	**AP1**,**AP3**,BLR,ETT,JAG,**LEAFY**, PI,RGA,SEP3
AT1G69120	AP1	12	**AG**,**AP1**,AP2,**AP3**,BLR,ETT, FLM,JAG,**LEAFY**,PI,SEP3,SVP
AT3G54340	AP3	13	**AG**,**AP1**,**AP3**,BLR,ETT,FLC, FLM,JAG,PI,RGA,SEP3,SOC1,SVP
AT4G18960	AG	7	**AG**,**AP3**,BLR,ETT,**LEAFY**,PI, SEP3

Table 6.1: List of the TFs of the 4 genes of interest established by [Chen et al., 2018]

Genes of the list in Table 6.1 are our references and we compare the set of candidates TFs obtained by the studied statistical methods with respect to them.

6.2 The transcriptomic database of *Arabidospis thaliana*

To get data on all genes simultaneously, we turn towards the *DNA microarray* technique [Lenoir and Giannella, 2006]. Technical details of the data extraction is proposed in Appendix 6.7. Here, we propose a description of the real transcriptomic dataset on which the statistical methods are applied.

The obtained intensity values are the ratio, at a given time, of the quantity of *transcripts* (mRNA fragments) produced in a first condition (stress, mutation,...) over the quantity of *transcripts* produced in the reference control of all genes simultaneously.

The public available transcriptomic databases of *Arabidospis thaliana* are abundant and are currently listed on the international website The Arabidopsis Information Resource (TAIR) [1]. In this thesis, we use the data grouped in the public database CATdb[2] [Gagnot et al., 2008]. They are produced by the transcriptomic plateform POPS[3]. The used *DNA microarray*, which covers almost the entire genome, was developed through the European CATMA (Complete Arabidopsis Transcriptome MicroArray) project [Crowe et al., 2003, Hilson et al., 2004].

In the database, the biological processes underway at the time of acquisition vary between projects. Hence, a gene expression profiling is available for each gene corresponding to the expression of the gene during these different biological steps. We assume that the expression of a target gene in an experiment is linked to expressions of all these TFs (over-expression for *activators* and sub-expression for *inhibitors*), and that conversely, if the target gene is not

[1] `https://www.arabidopsis.org/` for *Arabidopsis thaliana* [Swarbreck et al., 2007]

[2] `https://cat.opidor.fr/index.php/CATdb`

[3] `https://ips2.u-psud.fr/fr/plateformes/spomics-interatome-metabolome-transcriptome/pops-plateforme-transcriptomique.html`

expressed, its *transcription factors* have no reason to be expressed as well. Hence, under these assumption, there are therefore similarities in the expression profiles when the genes interact and the expression profiles study is adequate to identify interactions between genes, especially interactions between TFs for a target gene. Moreover, the data are from 13 different organs of *Arabidospis thaliana* (root, flower, leaf, stem,$\cdots$) and the first condition during data acquisition varies between projects. In particular, *Arabidospis thaliana* has been stressed in different ways by biotic stress (fungus, insect, bacteria,$\cdots$) and abiotic (heat, frost, drought, humidity,$\cdots$) or has undergone various mutations. Thus, the transcriptional activity of the whole genome from several experiments allows, in fine, to identify the *core transcriptome* involved in the mechanism of gene regulation and this, whatever the experimental condition.

In [Vasseur, 2017], the author pointed out that, through a differential analysis, ratio values on TFs vary according to the first condition (stress, mutation,...). Hence, determining the interactions between the TFs through the dynamics of response in the first condition is relevant. Moreover, there is information between TFs and target genes to be extracted from these data since Y.Vasseur showed that in average, expression levels of the TFs are comparable to those of non TFs. These data are candidates to recover the set of the TFs of a target gene whatever the first experimental condition.

The DNA of *Arabidopsis thaliana* is composed of 25498 genes including 2210 TFs grouped into 79 structural families according to the location of protein binding on the *promoter area* of the genes [Castrillo et al., 2011]. In our dataset, we have at hand a matrix in which each entry is a log-ratio of the gene expression between specific condition (stress, mutation) and the reference control. A value close to 0 means no variation of the gene expression; a value significantly larger than 0 (resp. smaller than 0) means over-expression (resp. sub-expression) in the specific condition with respect to the reference control. Each row of the matrix corresponds to an conditional experiment and each TF to a gene. These data were pre-processed and preliminary studied: 4789 genes are removed from the database because of anomalies during the extraction experiments. So, there are only biological replicates and no technical ones, so experiments can be considered as independent from each other). So, it remains 2215 independent conditional experiments of 19844 genes. Among them, 1887 are TFs. Figure 6.1 is an extraction of the table of the values of the matrix entries.

	AT4G17695	AT5G16470	AT3G27970	AT5G65670	AT3G07610	AT4G04580
X455	0.471	1.049	0.201	-0.900	-0.159	NA
X1338	-0.060	0.481	0.169	-1.081	-0.112	NA
X1339	0.009	-0.035	0.064	0.591	-0.093	NA
X1340	-0.150	-0.088	0.087	-0.943	-0.143	NA
X1341	0.030	0.341	0.078	-0.512	-0.016	NA
X1342	0.076	0.113	-0.033	0.164	0.181	NA

Figure 6.1: Extraction of the available data table. Each entry is the logarithm of the ratio of the gene expression between specific condition (stress, mutation) and the reference control.

6.3 Pre-processing and analyses of the available dataset

Theoretical results of statistical methods considered in this thesis are expressed with independent observations, Gaussian distribution and under the linear regression model. For the first aspect, there are only biological replicates and no technical ones in the dataset. So observations can be considered independent from each other. The last two aspects are more questionable for real data. This section is devoted to evaluate how far the data deviates from these properties and how to apply pre-processing.

In the following, the pair (Y, X) corresponds to the response variable Y modeling a target gene (LEAFY, AP1, AP3 or AG) and the matrix X is the matrix of all the TFs except the considered target gene. X is of size $n \times (p-1)$, with $n = 2215$ and $p = 1887$.

6.3.1 Missing values

Localization of missing values: The table of the values of the matrix entries contains many missing values. They are mainly caused by problems encountered during the production of the *DNA microarrays* and during the hybridization process. The percentage of missing data is 6%.

To deal with the missing values, we first study their distribution in the table. More precisely, we study the number of missing values per experimental condition and per TF. In Figure 6.2,

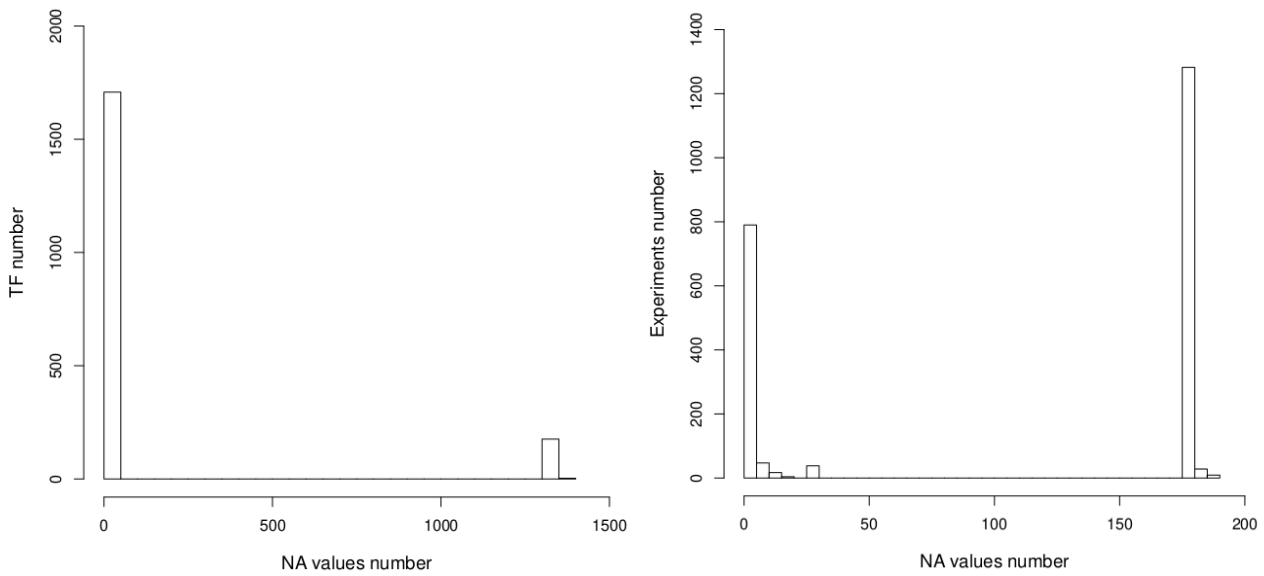


Figure 6.2: Histograms of the missing values distribution per TF (left) and per conditional experiment (right)

the missing value distributions per TF and per conditional experiment are represented via two histograms. Many TFs have few missing conditional experiments, while a peak is seen around

1400 missing conditional experiments (among 2215): this concerns about 200 TFs. At the scale of conditional experiments, a peak is observed around 180 TFs (among 1887): this concerns more than 1300 conditional experiments. The remaining missing values are concentrated on a few other conditional experiments. Thus, in conclusion, the missing values do not seem to be *Missing At Random (MAR)* but concentrated on a few TFs and among them, on a few conditional experiments.

Treatment of missing values: In-depth work on the treatment of missing data has not been done in this thesis. The first idea is the total removal of TFs and conditional experiments with at least one missing value. However, there are 412 TFs with at least one missing conditional experiment, including 179 ones with over 1300 missing conditional experiments. Removing these TFs may eliminate relevant TFs from the dataset. Regarding conditional experiments, 1513 have at least one missing TF data, including 1319 ones with more than 170 missing data. Removing these conditional experiments would remove two-thirds of the data. These two proposals do not seem a good idea and we propose a less brutal one.

Here are the steps of our missing data treatment process:

1. For each of the four pairs (Y, X), where Y models LEAFY, AP1, AP3 or AG, identify the conditional experiments of missing values on Y, remove them on Y and X.
2. Delete the remaining 179 TFs with more than 1300 missing expression data.
3. Remove the rest of the conditional experiments where missing values are still present.

The first step concerns only $Y = \text{AG}$ with only one missing value. At step 3, there are only 271 expression data left with at least one missing value for $Y = \text{AG}$ and 272 for the three others target genes.
At the end of the process, the final X matrix is identical for the four target genes and is of size 1708×1943. For the sequel, $n = 1708$ and $p = 1943$. So, $p = 1943$ is large and exceeds $n = 1708$ which puts the problem in a high-dimensional context. Almost 15% of the initial dataset are removed. Some relevant TFs may have been deleted but none of the four genes of interest is removed. The gene FLM is removed and is, according to the list of TFs in Table 6.1, a serious candidate in the regulation of for both AP1 and AP3. Lastly, the values $2k\log(\frac{p}{k})$ are always smaller than n for all $k \leq p$ avoiding the ultra high-dimensional setting [Verzelen et al., 2012] for all possible subsets of variables among the p ones.

6.3.2 Gaussian distribution

For each gene, the data table contains $n = 1708$ independent biological replicates under different experimental conditions. More precisely, each value is the log-ratio of the gene expression between two conditions: a positive value means an over-expression of the gene in the first condition in comparison to the second one; a negative value implies an under-expression of the gene in the first condition in comparison to the second one; a value close to 0 means a similar expression of the gene between the two conditions.

Modeling: The logarithm transformation allows to get unimodal, centered, symmetric and fine tailed distributions. Values are real and discrete but thanks to the *central limit theorem*, the Gaussian distribution modeling can be justified. We observe a zero-values inflation (see Figure 6.3).

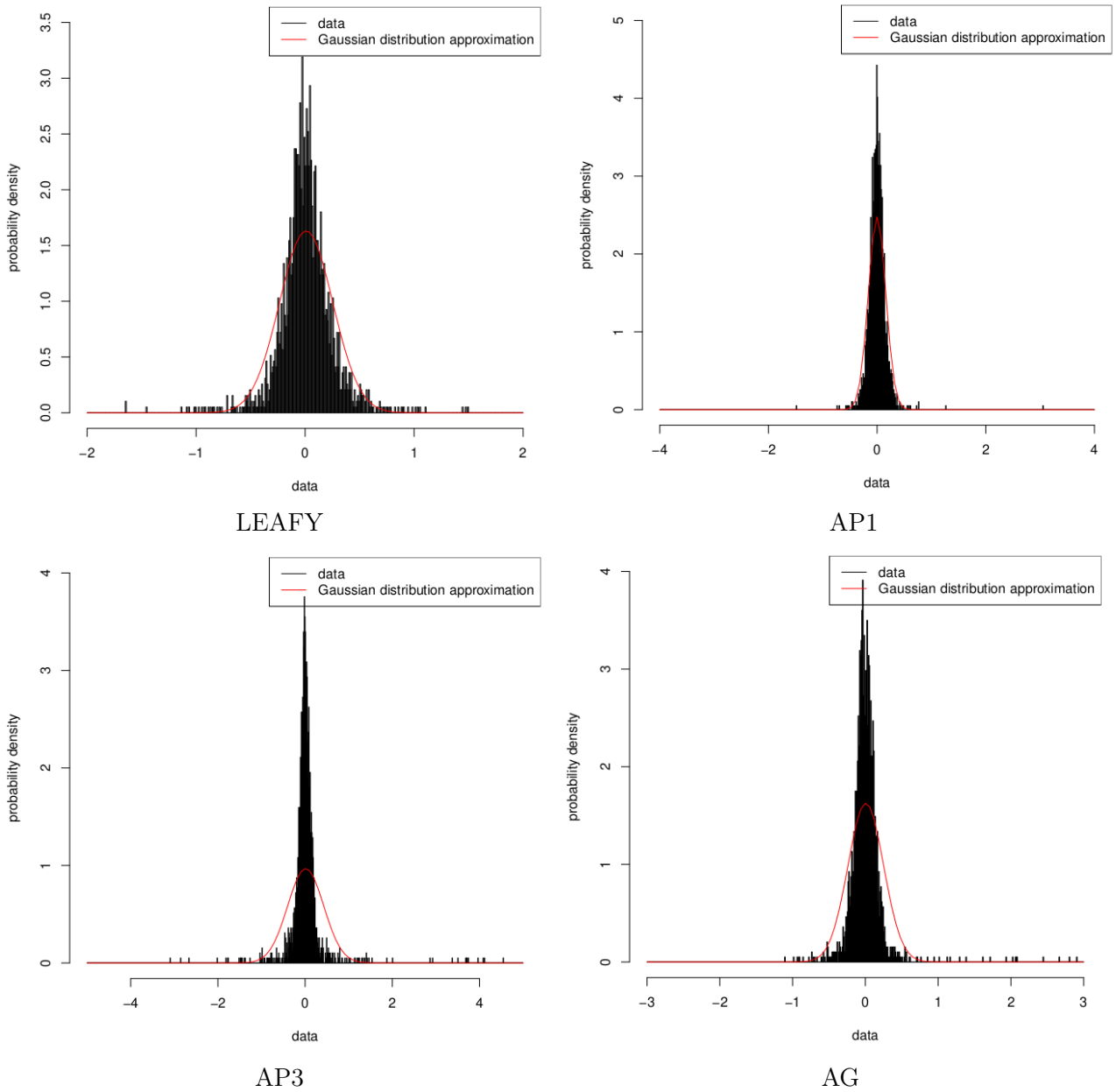


Figure 6.3: Black: Histograms of probability densities of the n available values for LEAFY, AP1, AP3 and AG. Red: Gaussian approximation by using the empirical estimation of the mean and the standard deviation.

For a more detailed study, we propose to apply the *Kolmogorov Smirnov test* on each TF

independently: after normalization of its expression vector, the cumulative function is tested against the standard Gaussian one. In the left Figure 6.4, distribution of the p-values obtained

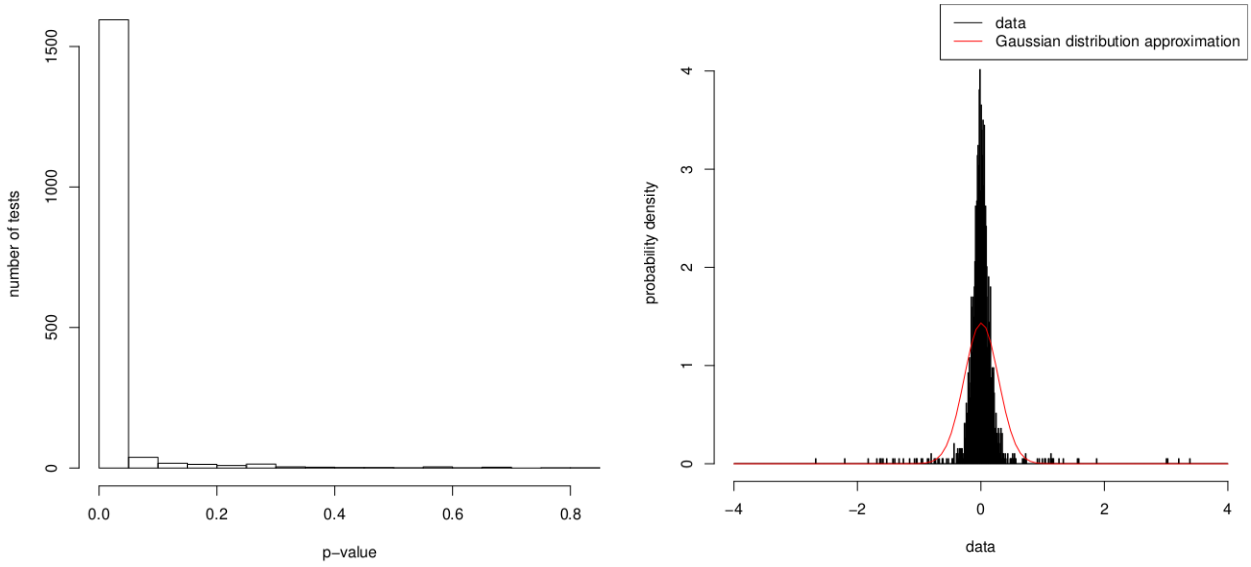


Figure 6.4: Left: Distribution of the p-values from the Kolmogorov Smirnov tests on the p TFs. Right: Histograms of the probability density of the n available data for the gene having the smallest Kolmogorov-Smirnov test p-value (black) and the Gaussian approximation by using the empirical estimation of the mean and the standard deviation (right).

by Kolmogorov Smirnov tests on the p TFs is represented. A large pick is observed around 0. More precisely, there are 1596 p-values smaller than 0.05. However, we decide to keep the Gaussian distribution modeling even if almost all tests are significantly rejected with very small p-values. The main reason is that distributions of the TFs have lighter tails than the standard Gaussian one leading to faster exponential decreases far from 0: real data distributions are more concentrated than the Gaussian one. So, the model selection theory would still be valid on the real data since it is mainly based on concentration inequalities.

Data transformation: The Gaussian distribution does not coincide perfectly with the real data distributions but we decide not to change them by some transformations. Indeed, the objective of this thesis is not to propose models to minimize the approximation error of the model with respect to the available data but to test and compare some statistical methods. So, we work with unprocessed log-ratio data to avoid the loss of biological information or potential correlations between the TFs.

6.3.3 Multivariate linear regression model

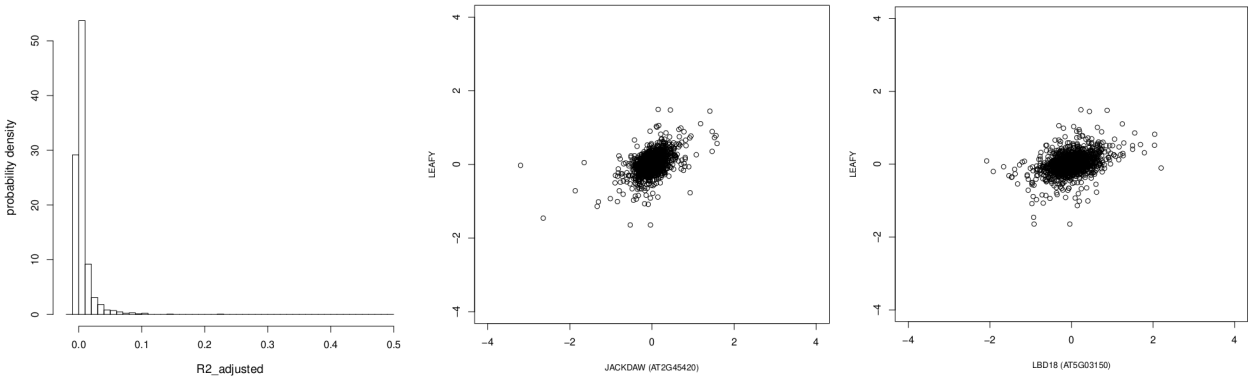
The *adjusted-R*2 evaluates the affine behavior between two vectors a and b and is defined by:

$$1 - \frac{\frac{\sum_{i=1}^{n}(a_i - \hat{\gamma} b_i)^2}{n-2}}{\frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n-1}},$$

where $\bar{a}$ is the mean of the vector a and $\hat{\gamma}$ is the multiplicative coefficient appearing in the linear regression of a against b. The adjusted-R2 value is the variability proportion of values of a explained by b thought a linear regression. The closer the value to 1, the better the linear adjustment is.

The logarithm transformation allows getting an additive model. In the left of Figure 6.5, the distributions of the adjusted-R2 between LEAFY and each of the other TFs are represented. We observe that many values are close to 0.02 (very small). The largest values are obtained for the JACKDAW (AT2G45420) and LBD18 (AT5G03150) TFs with small values 0.23 and 0.14. Hence, there is no Pearson correlation between LEAFY and any gene. Clouds of data between LEAFY and each of the two TFs show a slightly linear link. It suggests the presence of linear structures but not by pairs: we search for indirect correlations, called partial correlation (see Subsection 1.1.2.1 for more details). Hence, the multivariate linear regression model is justified.
The partial correlation is a symmetric statistical quantity. Therefore, a partial correlation between two TFs reflects the presence of a regulation link between them, but without any additional information, the regulation will not be oriented and it will not be possible to distinguish the target gene from its TF. To add the orientation, other statistical strategies have to be used on, for instance, on knock-out data using or on temporal data via Bayesian methods [Chen et al., 2006].

For the sequel, each of the genes of interest successively corresponds to the target gene and is considered as the response variable of the linear regression model defined in 5.1. The model considered is the Gaussian linear regression one, defined in (1.1). The non-zero coefficients of β^* correspond to the set of the TFs of the target gene. The non-zero coefficients of any estimator $\hat{\beta}$ of β^* give a set of candidates TFs. The main goal of this chapter is to know if the tested statistical methods recover the candidates TFs of Table 6.1 obtained by biological knowledge. The sparsity assumption is biologically justified since, according to biological knowledge, the number of TFs involved in regulating one target gene is small compared to the total number of TFs.



LEAFY | LEAFY and JACKDAW | LEAFY and LBD18

Figure 6.5: Left: Distribution of the adjusted-R2 values between LEAFY and each of the other TFs. Middle and right: Point clouds between the LEAFY and JACKDAW (middle) or LBD18 (right) data.

6.4 Evaluated statistical methods

In this chapter, we apply some of the statistical methods studied in Chapter 3, additionally to our 16 proposed algorithms described in Chapter 5. It is biologically known that TFs are correlated to each other. So, to select methods from Chapter 3 we turn to conclusions of recommendations part for *cluster* and *scale-free* settings (Subsection 3.5). More precisely, model collections are generated from the Elastic-Net criterion minimization and with the LARS algorithm. The penalty eBIC is one of the best statistical methods among those considered in Chapter 3 to control the MSE and the trade-off between recall and specificity; LinSelect and the knockoffs method are one of the best statistical methods to control both the MSE and the FDR; Bolasso (when samples are first generated) and Tigress are included in the study since they get the smallest FDR; our algorithms of Chapter 5 control the best the MSE. We do not include our algorithm controlling the trade-off between the predictive risk and the FDR (Subsection 4.4.2 of Chapter 4). The reason is that the data-dependent calibration algorithm is based on theoretical results with assumptions $\mathcal{M}$ deterministic, ordered variable selection and $m^* \in \mathcal{M}$. Any of these assumptions is satisfied by the model collections generated on the real data.

Implementation parameters We use almost the same parameters as those used in Chapter 3. For the LARS algorithm, we use the function enet of the R package elasticnet (version 1.1.1). The maximal number of steps to define the grid size is the default value $50 \times \min(p, n-1)$. We set $\alpha = 0.5$ for Elastic-Net regularization. The penalty eBIC is implemented with δ equals 1. Concerning LinSelect, it is implemented in the function tuneLasso of the R package LINselect (version 1.1.3). The data-driven penalties are calculated by using the function capushe of the R package capushe (version 1.1.1) for the penalty *Capus_2.5* and parameters are set to the default values except for the minimum percentage of points for the plateau selection

set to 0.05 (instead of 0.1 in Chapter 3). Only 50 resamples are used for Bolasso because of the computational cost (instead of 100 in Chapter 3) and a variable is selected when its occurrence frequency is higher than 0.8. To apply Tigress, we use the function tigress of the R package tigress (version 0.1.0) with 50 steps for the LARS algorithm. A randomized Lasso principle is processed for Bolasso and Tigress with weights drawn from a uniform distribution between 0.7 and 1. For the knockoffs method, we use the function knockoff.filter with option create.second_order of the R package knockoff (version 0.3.2), we calculate the W_j with the function stat.lasso_lambdasmax and set the false discovery rate to 0.1.

6.5 Results

Tables 6.2- 6.5 summarize the set of selected TFs for respectively target genes LEAFY, AP1, AP3 and AG obtained by the evaluated statistical methods listed in Section 6.4. Tables 6.6-6.9 summarize the number of each TF of the list in Table 6.1 in the model collection. Numbers in brackets mean that the TF is selected by the statistical method and is the number appearing in the corresponding model collection; lastly "-" means that the TF is not selected. When the model collection is *path_Bolasso_all_models*, TFs are not ordered since all the models of all the collections are considered during the model selection step. Hence, "YES" means that the statistical method selects the TF. The denominations *path*, *path_Bolasso_LS_each_resample*, *path_Bolasso_LS_fulldata path_Bolasso_all_models*, *Capus_2.5*, *2D_Stop_End*, *2D_Stop_Plat*, *2D_Full_End* and *2D_Full_Plat* are the same as in Chapter 5. The denomination *path_bis* in Tables 6.6-6.9 denotes the model collection used for eBIC, LinSelect, Bolasso, Tigress and the knockoffs method. It slightly differs from *path* since pre-processing on the model collection is applied before the model selection process by *Capus_2.5* (see Chapter 5 for more details). According to the list of Table 6.1, we recall that LAEFY is removed from the X matrix when it is the target gene. Moreover, FLM no longer appears in the dataset after the pre-processing step (Section 6.3).

According to Tables 6.2- 6.5, the first striking remark is that among the candidates TFs of Table 6.1, there are more non-selected TFs than selected ones.
One analysis per TF allows us to observe that no TF is never selected. Moreover, the different statistical methods often select the same TFs: AP1 and JAG for LEAFY, LEAFY, AP3, BLR and PI for AP1, AP1, AG, PI and SEP3 for AP3, AP3, PI and SEP3 for AG. The others are either never selected or selected by only one method. In this last case, its number in the model collection is significantly higher than for the often selected TFs.
One analysis per statistical method allows to observe that both Bolasso and the knockoffs method never select a TF appearing in the list of Table 6.1. Concerning the knockoffs method, the selected TFs set is always empty. Concerning Bolasso, the selected TFs sets are not empty and of size 34, 36 or 38 but never contain the TFs in the list of Table 6.1. A surprising observation is that LinSelect selects more TFs than eBIC when AP3 is the target gene, while according to Chapter 3, LinSelect is more conservative than eBIC. Concerning the 2D-slope heuristics

algorithms developed in chapter 5, the selected TFs sets are all empty for AP1, AP3 and AG as target genes except when the model collection is *path*; some of them contain TFs when LEAFY is the target gene, but the set size does not exceed 45 except when the model collection is *path*. Concerning our 2D-slope heuristics algorithms, when the model collection is path, the size of the selected TFs set is large. All the existing methods get a reasonable selected TF set with a size smaller than 48. They select at most 0 over 8 TFs of the list in Table 6.1, 1 over 10, 4 over 11 and 3 over 6 for respectively LEAFY, AP1, AP3 and AG as target gene. Concerning the 2D-slope heuristics algorithms developed in chapter 5, methods with the model collection *path* select several or all the candidates TFs of Table 6.1 but the size of the selected TFs set is very large (until 1707). The size is mostly large for all the others when Y models LEAFY but null or equals 29 or 41 for the other target genes.
According to Tables 6.6-6.9, the higher the number, the later the apparition of the TF is in the model collection. When LEAFY is the target gene, all the TFs of Table 6.1 appear after the 125th. Thus any selected model of reasonable size cannot contain any TFs. Except for the model collection *path_Bolasso_LS_each_resample*, AP3 and AG appear early in the model collection (before the 22th) when the target gene is AP1. Concerning AP3, viewed as the target gene, AG, PI and SEP3 belong to the 10 first TFs appearing in all the model collections. Lastly, when AG is the target gene, AP3, PI and SEP3 belong to the 8 first TFs appearing in all the model collections. The model collection *path_Bolasso_LS_fulldata* seems to sort the best the TFs according to the list in Table 6.1, *path_bis* and *path* have the highest values.

In this paragraph, the conclusions of the analyses above are given according to the list in Table 6.1. Among this list of TFs, some of them are recovered by several methods: AP1 and JAG for LEAFY but these TFs appear late in model collections; PI appears early in model collections and is often selected to regulate AP1, AP3 and AG. It is also the case for AP3 and SEP3 to regulate AP1 and AG, and AP3 and AG respectively. Moreover, pairs of TFs can be identified: LEAFY and AP1, AG and AP3; but, as the partial correlation is a symmetric statistical quantity, any orientation of the regulation can be proposed. The comparison study on the model collections suggests that adding a resampling process in the model collection improves the order in which the TF appears in the model collection.

6.6 Conclusions

According to analyses of Section 6.5, the different statistical methods behave similarly: either the TF is selected and appears early in the model collections, or the TF is selected by only one type of method and appears late in the model collections. Selecting the same TFs is encouraging to validate the statistical model and approach. Moreover, several statistical methods have recovered some biologically known regulatory links in Table 6.1 and appear early in the model collections, particularly the reciprocal link between AP3 and AG. It may encourage further bench experiments to test other regulatory links in the selected TFs sets. However, there are more non-selected TFs than selected ones among the list. Hence, three conclusions can be

possible:

1. the biological knowledge is still sparse. The more relevant TFs to regulate the target genes are not among those in the list of Table 6.1.
2. all the statistic methods give bad results. The modeling is not adapted to the biological question.
3. all the statistic methods give bad results. The transcriptomic data is not adapted to the biological question.

Concerning conclusion 2., the modeling choices are of different scales: the missing values treatment, the Gaussian distribution choice, the data transformation through normalization or distribution modification, the multivariate linear regression modeling or the processed statistical methods. A perspective is to study each step. The pre-processing question is complex and sensitive since data have to be adapted to the statistical model and, simultaneously, be transformed as little as possible to avoid loss or modification of the information.
The fact that the knockoffs method systematically gives an empty selected TF set may be of interest. A reason may be at the model collection step: TFs may be disordered and any model satisfies a FDR smaller than 10%. An interpretation may also question the conclusion 1.
Conclusions 2. and 3. can be studied in parallel. Indeed, regulation links can not be partial correlation links between genes. In this case, either transcriptomic data can be adapted to answer the biological question but another statistical modeling has to be used, or a multivariate linear regression model can be adapted but processed with another type of dataset.

To conclude, this chapter does not finally answer the biological problem but proposes three possible solutions.

Table 6.2: Selected TFs for LEAFY

LEAFY	number of selected TFs	***AP1***	***AP3***	BLR	ETT	JAG	PI	RGA	SEP3
eBIC	39	-	-	-	-	-	-	-	-
LinSelect	38	-	-	-	-	-	-	-	-
Bolasso	34	-	-	-	-	-	-	-	-
Tigress	1	-	-	-	-	-	-	-	-
knockoffs	0	-	-	-	-	-	-	-	-
Capus_2.5	39	-	-	-	-	-	-	-	-
path:									
2D_Stop_End	277	(247)	-	-	-	-	-	-	-
2D_Stop_Plat	398	(247)	-	-	-	-	-	-	-
2D_Full_End	637	(247)	-	-	-	-	(626)	-	(447)
2D_Full_Plat	277	(247)	-	-	-	-	-	-	-
path_Bolasso_LS _each_resample:									
2D_Stop_End	534	(341)	-	-	-	(530)	-	-	-
2D_Stop_Plat	1	-	-	-	-	-	-	-	-
2D_Full_End	29	-	-	-	-	-	-	-	-
2D_Full_Plat	534	(341)	-	-	-	(530)	-	-	-
path_Bolasso _LS_fulldata:									
2D_Stop_End	584	-	-	-	(511)	-	-	-	-
2D_Stop_Plat	388	(173)	-	-	-	-	-	-	-
2D_Full_End	4	-	-	-	-	-	-	-	-
2D_Full_Plat	1	-	-	-	-	-	-	-	-
path_Bolasso _all_models:									
2D_Stop_End	45	-	-	-	-	-	-	-	-
2D_Stop_Plat	225	YES	-	-	-	YES	-	-	-
2D_Full_End	388	YES	-	-	-	-	-	-	-
2D_Full_Plat	326	YES	-	-	-	YES	-	-	-

Table 6.3: Selected TFs for AP1

AP1	number of selected TFs	***LEAFY***	***AP3***	***AG***	AP2	BLR	ETT	JAG	PI	SEP3	SVP
eBIC	4	-	(2)	-	-	-	-	-	-	-	-
LinSelect	4	-	(2)	-	-	-	-	-	-	-	-
Bolasso	38	-	-	-	-	-	-	-	-	-	-
Tigress	2	-	-	-	-	-	-	-	(188)	-	-
knockoffs	0	-	-	-	-	-	-	-	-	-	-
Capus_2.5	4	-	(3)	-	-	-	-	-	-	-	-
path:											
2D_Stop_End	448	(55)	(3)	(8)	-	-	-	-	(189)	-	-
2D_Stop_Plat	1707	(55)	(3)	(8)	(479)	(935)	(1018)	(575)	(189)	(1536)	(1407)
2D_Full_End	463	(55)	(3)	(8)	-	-	-	-	(189)	-	-
2D_Full_Plat	56	(55)	(3)	(8)	-	-	-	-	-	-	-
path_Bolasso_LS_each_resample:											
2D_Stop_End	0	-	-	-	-	-	-	-	-	-	-
2D_Stop_Plat	0	-	-	-	-	-	-	-	-	-	-
2D_Full_End	0	-	-	-	-	-	-	-	-	-	-
2D_Full_Plat	0	-	-	-	-	-	-	-	-	-	-
path_Bolasso_LS_fulldata:											
2D_Stop_End	0	-	-	-	-	-	-	-	-	-	-
2D_Stop_Plat	0	-	-	-	-	-	-	-	-	-	-
2D_Full_End	0	-	-	-	-	-	-	-	-	-	-
2D_Full_Plat	0	-	-	-	-	-	-	-	-	-	-
path_Bolasso_all_models:											
2D_Stop_End	0	-	-	-	-	-	-	-	-	-	-
2D_Stop_Plat	0	-	-	-	-	-	-	-	-	-	-
2D_Full_End	41	YES	-	-	-	YES	-	-	-	-	-
2D_Full_Plat	29	YES	-	-	-	YES	-	-	-	-	-

Table 6.4: Selected TFs for AP3

AP3	number of selected TFs	***AP1***	***AG***	BLR	ETT	FLC	JAG	PI	RGA	SEP3	SOC1	SVP
eBIC	8	-	-	-	-	-	-	(3)	-	(2)	-	-
LinSelect	44	(44)	(9)	-	-	-	-	(3)	-	(2)	-	-
Bolasso	34	-	-	-	-	-	-	-	-	-	-	-
Tigress	2	-	-	-	-	-	-	(3)	-	-	-	-
knockoffs	0	-	-	-	-	-	-	-	-	-	-	-
Capus_2.5	48	(44)	(10)	-	-	-	-	(4)	-	(3)	-	-
path:												
2D_Stop_End	473	(44)	(10)	-	-	-	-	(4)	-	(3)	-	-
2D_Stop_Plat	1707	(44)	(10)	(1518)	(871)	(1476)	(959)	(4)	(1109)	(3)	(1301)	(1382)
2D_Full_End	644	(44)	(10)	-	-	-	-	(4)	-	(3)	-	-
2D_Full_Plat	279	(44)	(10)	-	-	-	-	(4)	-	(3)	-	-
path_Bolasso_LS_each_resample:												
2D_Stop_End	0	-	-	-	-	-	-	-	-	-	-	-
2D_Stop_Plat	0	-	-	-	-	-	-	-	-	-	-	-
2D_Full_End	0	-	-	-	-	-	-	-	-	-	-	-
2D_Full_Plat	0	-	-	-	-	-	-	-	-	-	-	-
path_Bolasso_LS_fulldata:												
2D_Stop_End	0	-	-	-	-	-	-	-	-	-	-	-
2D_Stop_Plat	0	-	-	-	-	-	-	-	-	-	-	-
2D_Full_End	0	-	-	-	-	-	-	-	-	-	-	-
2D_Full_Plat	0	-	-	-	-	-	-	-	-	-	-	-
path_Bolasso_all_models:												
2D_Stop_End	0	-	-	-	-	-	-	-	-	-	-	-
2D_Stop_Plat	0	-	-	-	-	-	-	-	-	-	-	-
2D_Full_End	0	-	-	-	-	-	-	-	-	-	-	-
2D_Full_Plat	0	-	-	-	-	-	-	-	-	-	-	-

Table 6.5: Selected TFs for AG

AG	number of selected TFs	***LEAFY***	***AP3***	BLR	ETT	PI	SEP3
eBIC	11	-	(5)	-	-	(4)	(2)
LinSelect	2	-	-	-	-	-	(2)
Bolasso	36	-	-	-	-	-	-
Tigress	3	-	-	-	-	(4)	(2)
knockoffs	0	-	-	-	-	-	-
Capus_2.5	22	-	(6)	-	-	(5)	(3)
path:							
2D_Stop_End	1707	(1558)	(6)	(85)	(299)	(5)	(3)
2D_Stop_Plat	1707	(1558)	(6)	(85)	(299)	(5)	(3)
2D_Full_End	464	-	(6)	(85)	(299)	(5)	(3)
2D_Full_Plat	367	-	(6)	(85)	(299)	(5)	(3)
path_Bolasso_LS _each_resample:							
2D_Stop_End	0	-	-	-	-	-	-
2D_Stop_Plat	0	-	-	-	-	-	-
2D_Full_End	0	-	-	-	-	-	-
2D_Full_Plat	0	-	-	-	-	-	-
path_Bolasso _LS_fulldata:							
2D_Stop_End	0	-	-	-	-	-	-
2D_Stop_Plat	0	-	-	-	-	-	-
2D_Full_End	0	-	-	-	-	-	-
2D_Full_Plat	0	-	-	-	-	-	-
path_Bolasso _all_models:							
2D_Stop_End	0	-	-	-	-	-	-
2D_Stop_Plat	0	-	-	-	-	-	-
2D_Full_End	0	-	-	-	-	-	-
2D_Full_Plat	0	-	-	-	-	-	-

Table 6.6: Order of the LEAFY selected TFs in paths

LEAFY	***AP1***	***AP3***	BLR	ETT	JAG	PI	RGA	SEP3
path_bis	248	1390	1351	1418	790	629	1566	450
path	247	1386	1347	1414	786	626	1562	447
path_Bolasso_LS_each_resample	341	244	262	284	530	529	898	343
path_Bolasso_LS_fulldata	173	724	125	511	210	772	298	298

Table 6.7: Order of the AP1 selected TFs in paths

AP1	***LEAFY***	***AP3***	***AG***	AP2	BLR	ETT	JAG	PI	SEP3	SVP
path_bis	54	2	7	480	940	1023	576	188	1542	1413
path	55	3	8	479	935	1018	575	189	1536	1407
path_Bolasso_LS_each_resample	808	94	54	808	868	821	430	361	808	833
path_Bolasso_LS_fulldata	65	9	21	788	669	172	570	65	467	889

Table 6.8: Order of the AP3 selected TFs in paths

AP3	***AP1***	***AG***	BLR	ETT	FLC	JAG	PI	RGA	SEP3	SOC1	SVP
path_bis	44	9	1524	875	1482	963	3	1113	2	1307	1388
path	44	10	1518	871	1476	959	4	1109	3	1301	1382
path_Bolasso_LS_each_resample	52	6	305	189	789	481	5	673	4	693	478
path_Bolasso_LS_fulldata	60	10	876	1043	1043	992	5	1186	3	992	185

Table 6.9: Order of the AG selected TFs in paths

AG	***LEAFY***	***AP3***	BLR	ETT	PI	SEP3
path_bis	1558	5	84	298	4	2
path	1558	6	85	299	5	3
path_Bolasso_LS_each_resample	600	3	134	454	7	4
path_Bolasso_LS_fulldata	269	8	23	239	5	3

6.7 Appendix: The *DNA microarray* technique

Appearing at the end of the 1990s, *DNA microarrays* on glass slides allow the study of the expression of thousands of genes simultaneously. For each gene, the ratio of its expressions is acquired between two conditions at a given time. Usually, the first condition is a stressed environment (biotic or abiotic) or a mutation and the second one is a reference control.

The *DNA microarray* technique relies mainly on the principle of complementarity pairing revealed during the DNA *transcription*. Many DNA fragments are first deposited on the *microarray*, and the *transcripts* (the mRNA fragments) of the collected sample are then attached to the *microarray* at their complement. Note that it is therefore necessary that the sequencing of the organism's DNA be processed beforehand, which is the case with *Arabidospis thaliana*. More precisely, the *DNA microarray* principle [Lenoir and Giannella, 2006] is split into four steps:
Production of *DNA microarrays*: on one side, thousands of single-stranded DNA fragments are deposited on the glass slide, in an organized way, to form several points, called spot. Some DNA fragments can be the replication of the same gene but each spot corresponds to one specific gene. They are then amplified by PCR (Polymerase Chain Reaction).
Preparation of targets: On the other side, *transcripts* (mRNA fragments) are collected from the organism, in two different conditions and at a given time. A reverse *transcription* is performed to obtain complementary fragments, artificially synthesized, to those on the *microarray*. These fragments are called cDNA fragments. The cDNA fragments from the first condition are marked by the fluorescent tracer Cy3 (green color), while the ones from the reference control are marked by the fluorescent tracer Cy5 (red color), before being mixed together.
Hybridization: The mixed cDNA fragments are deposited on the *DNA microarray*. Each strand of cDNA is paired by complementarity to a strand of DNA, to form the double helix. The unpaired cDNA are then eliminated.
Reading the results: The fluorescence of each spot is measured with a scanner. This allows to recover the cDNA amount from the two distinct conditions that has been matched to the DNA fragments of each spot. Two images are generated by the scanner, one for each condition, and a dot on the image represents a spot. These are in grayscale and represent the fluorescent signal intensity of each spot. The grayscale of the first image is replaced by a greenscale image; the second one is replaced by a redscale image; and both images are overlapping given an artificial image from green (over-representation of the first condition) to red (over-representation of the reference control). The yellow color represents an equal proportion of the two conditions on the corresponding spot. Due to the large amount of DNA fragments (approximately thousands of fragments) present upstream on the *microarray*, the image shows a wide range of fluorescence intensity values.

Nowadays, the *DNA microarray* techniques are outdated and replaced by high-throughput sequencing technologies like the RNA sequencing (RNA-seq) [Morin et al., 2008,Chu and Corey, 2012]. However, with these extraction methods, the data produced are count data (natural integers) and so that are modeled by a discrete distribution. This is not the case when they are extracted from *DNA microarray* principles where data are real numbers and lead to distribution

close to a continuous one. The modeling with a continuous distribution is so more appropriated and it is more satisfying for our statistical procedure.

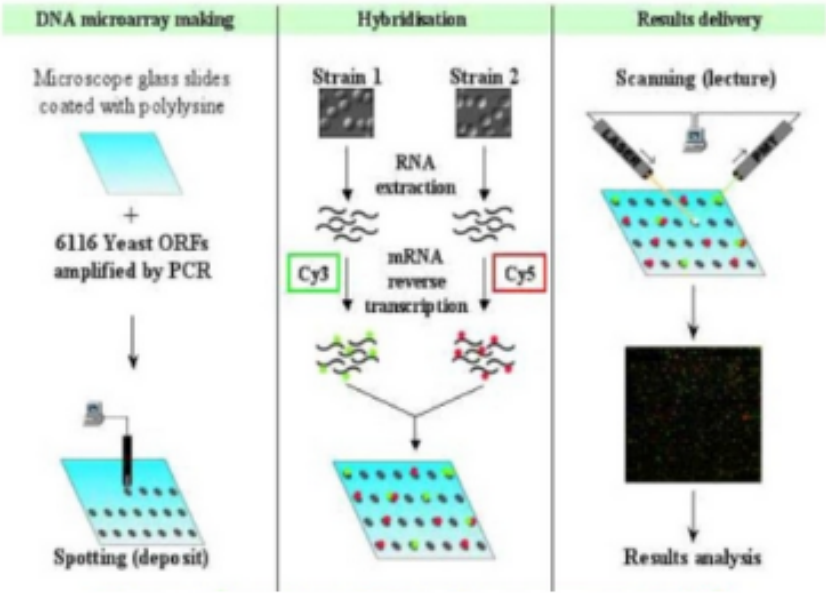


Figure 6.6: Illustration of the *DNA microarray* principle [4]

[4] https://universe84a.com/dna-microarray/

Conclusion et Perspectives

Cette thèse propose de nouvelles méthodologies statistiques, toutes construites à partir de la problématique biologique soulevée à l'amorce de ces travaux. Plus précisément, cette thèse traite la question de sélection de variables lorsque les données sont distribuées selon une loi gaussienne, en adéquation avec un modèle de régression linéaire et inscrites dans un contexte de grande dimension. Nous proposons de combiner la sélection de modèle qui est adaptée à la sélection de variables mais qui a coût computationnel trop élevé pour la grande dimension avec l'optimisation convexe qui est adaptée à la grande dimension mais qui s'éloigne de la problématique de la sélection de variables. Nous proposons deux nouvelles fonctions de pénalité. Toutes deux dépendent d'hyperparamètres dont la calibration est réalisée uniquement sur le jeu de données disponible. La première calibration combine l'approche prédictive et l'approche FDR (chapitre 4) ; la seconde généralise le principe de l'estimation des pentes à deux constantes et pour une collection de modèles aléatoire (chapitre 5).

Ci dessous sont proposées des perspectives de travail, au delà de celles mentionnées à la fin de chacun des chapitres précédents.

- La forme de la pénalité :
Cette thèse est entièrement consacrée à la fonction de pénalité :

$$\mathrm{pen}_{\mathrm{opt}}(m) = K\left(C_1(\sigma^2)\frac{D_m}{n} + C_2(\sigma^2)\frac{\log(\binom{p}{D_m})}{n}\right), \tag{6.1}$$

où une calibration de $K > 1$ est proposée dans le chapitre 4 et une calibration du couple $\left(C_1(\sigma^2), C_2(\sigma^2)\right)$ est proposée dans le chapitre 5. Bien que cette forme précise de fonction est issue des résultats théoriques (Théorème 1. de 5.7) pour contrôler le risque prédictif d'une manière optimale et non-asymptotique, elle s'appuie sur de nombreuses approximations avant d'obtenir la forme (6.1), comme par exemple :

- sa forme est une simplification de la forme théorique exacte, dont l'erreur d'approximation est contrôlée par un terme négligeable via des inégalités de concentration
- l'inégalité $2ab \leq a^2 + b^2$ est utilisée pour simplifier les calculs.

Ainsi, d'autres formes de pénalités peuvent être envisagées alternativement.
Par exemple, le chapitre 5 généralise la méthode de l'estimation des pentes à la calibration de deux hyperparamètres simultanément. Cette estimation est d'autant plus difficile si $|\beta^*|$ est grand : l'aspect affine attendu théoriquement n'est présent que sur quelques modèles de la collection, ce qui rend difficile l'estimation des deux constantes. Une idée serait de tenir compte de $|\beta^*|$ dans la forme de pénalité. Trancher entre "$|\beta^*|$" petit et "$|\beta^*|$" grand peut être inclus au sein des algorithmes de l'estimation des pentes :

- si l'aspect linéaire est visible sur beaucoup de modèles de la collection, alors la zone de sur-apprentissage est grande et $|\beta^*|$ est petit. Deux hyperparamètres peuvent être raisonnablement estimés et

$$\mathrm{pen}_{\mathrm{shape}}(m) = \left(\frac{D_m}{n}, \frac{\log(\binom{p}{D_m})}{n} \right)$$

  comme forme de pénalité peut être considérée

- si l'aspect linéaire est visible sur peu de modèles de la collection, alors la zone de sur-apprentissage est petite et $|\beta^*|$ est grand. Un seul hyperparamètre peut être raisonnablement estimé avec comme forme de pénalité possible :

$$\mathrm{pen}_{\mathrm{shape}}(m) = \left(\frac{D_m}{n} \right) \quad \text{ou} \quad \mathrm{pen}_{\mathrm{shape}}(m) = \left(\frac{D_m}{n} \left(2.5 + \log\left(\binom{p}{D_m} \right) \right) \right)$$

Un autre exemple, inspiré des travaux du chapitre 4 est d'introduire l'approche FDR au sein de la forme de la pénalité, en plus de la calibration de K. Par exemple, chaque modèle pourrait être pénalisé par la somme des p-valeurs des variables qu'il contient : plus celles-ci sont grandes, plus la pénalité est forte ; une procédure de tests multiples pourrait être réalisée sur chaque modèle pour mesurer son FDR, dans la même lignée des travaux de [Abramovich et al., 2006].

- La sélection du modèle :
Le chapitre 4 met en place un algorithme pour sélectionner un modèle de la collection. Le problème d'optimisation est la minimisation du risque prédictif sous contrainte d'un contrôle du FDR fixé par l'utilisateur. Ce critère peut être changé en :

- la minimisation du FDR sous contrainte d'un contrôle du risque prédictif fixé par l'utilisateur

- la minimisation d'un compromis entre les deux métriques :

$$\widehat{K}(\mu) = \arg\min_{K>1} \{ \widehat{RP}(\hat{m}(K)) + \mu \widehat{FDR}(\hat{m}(K)) \},$$

  où $\widehat{RP}$ et $\widehat{FDR}$ estiment respectivement le RP et le FDR des modèles de la collection et où $\mu > 0$ est un nouvel hyperparamètre pondérant le compromis des deux fonctions de coût

- une procédure d'agrégation de modèles : procéder à l'approche prédictive et l'approche FDR individuellement et agréger les ensembles de variables sélectionnées.

- La collection de modèle :
Dans cette thèse, la collection de modèles est construite via un chemin de régularisation. Ce dernier est obtenu par la minimisation d'un critère convexe (Lasso, Elastic-Net,...). Cependant, comme le souligne [Su et al., 2017], pour des données gaussiennes de grande dimension, la première variable non active arrive très tôt dans le chemin de régularisation. De plus, les auteurs montrent qu'il est impossible de contrôler simultanément et correctement l'erreur de première espèce et l'erreur de seconde espèce le long du chemin. La raison principale est que le chemin de régularisation est construit via une approche prédictive.
En travail complémentaire, nous avons également observé ce phénomène sur les données simulées utilisées dans cette thèse. Par exemple, l'algorithme du chapitre 4 a été appliqué sur les mêmes données mais à partir d'une collection de modèle aléatoire, issue d'un chemin de régularisation Lasso. Le FDR de chaque modèle y est très élevé et donc, pour toute valeur de $K \in [1, 10]$, le FDR de $\widehat{m}(K)$ l'est aussi. En fait, des variables non actives arrivent très tôt dans le chemin de régularisation et certaines variables actives apparaissent très tard.
Un travail plus poussé a montré qu'introduire une procédure de ré-échantillonnage et ranger les variables selon leur fréquence d'apparition dans les chemins de régularisations (comme Bolasso) permet de mieux ordonner les variables de sorte que les variables actives soient les premières et les non actives les dernières. Cependant, l'ordre obtenu n'est pas parfait et il existe souvent un nombre non négligeable de variables non actives avec une fréquence d'apparition plus élevées que certaines variables actives : ces variables non actives sont donc robustes au ré-échantillonage. L'introduction d'une copie neutre (la méthode des knockoffs) et l'utilisation des statistiques W_j pour ranger les variables ne semble pas améliorer l'ordre des variables dans un chemin de régularisation type Lasso.
Dans cette thèse, seuls les estimateurs Lasso et Elastic-Net avec $\alpha = 0.5$ ont été utilisés. Une autre valeur de α ou d'autres estimateurs sont à explorer. Par exemple, sur nos données biologiques, le résultat d'un test de Bonferroni est disponible sur chaque observation. Celui-ci teste si le log-ratio entre les deux expressions du gène est significativement différent de 0 ou non. Les p-valeurs sous-jacentes peuvent être utilisées pour classer les gènes dans différents groupes de gènes corrélés puisqu'il est attendu que des gènes corrélés aient des profils d'expression similaires. Dans ce cas, le *Groupe Lasso* ou le *Lasso adaptatif* peuvent être un bon choix d'estimateur.
Le chapitre 4 propose de combiner l'approche prédictive et l'approche FDR lors de l'étape de la sélection du modèle. Il est tout à fait envisageable de considérer ces deux approches dès l'étape de la construction de la collection de modèles. Si la sélection du modèle est effectuée via une approche prédictive (pénalisation des moindres carrés), la collection de modèles peut-être créée via l'approche des tests multiples en s'inspirant de [Benjamini and Hochberg, 2000] : minimisation en α de la quantité $\sup_{1 \leq i \leq p} \left(p_{(i)} \leq \alpha\right)$ pénalisée par une fonction dépendante de α, similairement à l'estimateur Lasso où les moindres carrés sont pénalisés par la norme ℓ_1; ou en s'inspirant de [Genovese and Wasserman, 2002, Genovese et al., 2004] : minimisation de

différents compromis entre le FDR et le FNR ; ou en s'inspirant des travaux de [Benjamini and Hochberg, 1995,Storey, 2002] : les modèles seraient construits en contrôlant le FDR, éventuellement estimé, dont le contrôle serait de moins en moins fort au fur et à mesure de la construction de la collection. Cependant, le désavantage de ces collections pourrait être que les modèles ne satisfassent plus de performances prédictives. Ainsi, une solution serait de contrôler un compromis entre le RP et le FDR dès l'étape de la construction de la collection de modèles, en utilisant ce qui a déjà été proposé dans [Abramovich et al., 2006] et [Bogdan et al., 2013], ou en proposant de nouvelles approches comme par exemple la double minimisation sous contraintes simultanées de l'équation Lasso et du FDR contrôlé.

- La question biologique et le modèle statistique :
Les chapitres 3 et 6 ont montré que le modèle statistique proposé n'était pas tout à fait adapté aux données FRANK et aux données transcriptomiques réelles. L'hypothèse gaussienne n'est pas totalement vérifiée ni le comportement linéaire. Une perspective pour apporter une nouvelle conclusion biologique est d'évaluer l'erreur d'approximation du modèle vis-à-vis du jeu de données réel ; une autre serait de modifier le modèle pour diminuer cette erreur. D'un point de vue statistique, cela revient à adapter nos procédures sur le modèle linéaire généralisé où la distribution n'est plus nécessairement gaussienne, où l'homoscédasticité ne serait plus nécessairement vérifiée, où X ne serait plus nécessairement fixe mais aléatoire, voire même adapter nos procédures sur un modèle non linéaire.
Le chapitre 6 ne fournit que des listes de *facteurs de transcription* candidats à être acteur dans le mécanisme de régulation d'un gène cible. Pour améliorer l'information apportée par la statistique, une perspective serait d'ajouter une valeur sur chaque FT sélectionné qui traduirait la "croyance" en ce FT pour le gène cible correspondant.

Bibliography

[Abramovich et al., 2006] Abramovich, F., Benjamini, Y., Donoho, D. L., Johnstone, I. M., et al. (2006). Adapting to unknown sparsity by controlling the false discovery rate. The Annals of Statistics, 34(2):584–653.

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In Proc. 2nd Int. Symp. on Information Theory, pages 267–281.

[Allen, 1974] Allen, D. (1974). The relationship between variable selection and data augmentation and slow feature analysis. Technometrics, 16:125–127.

[Arlot, 2011] Arlot, S. (2011). Sélection de modèles et sélection d'estimateurs pour l'apprentissage statistique (cours peccot) premier cours: Apprentissage statistique et sélection d'estimateurs.

[Arlot, 2019] Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. arXiv preprint arXiv:1901.07277.

[Arlot and Celisse, 2010] Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4:40–79.

[Arlot et al., 2019] Arlot, S., Celisse, A., and Harchaoui, Z. (2019). A kernel multiple change-point algorithm via model selection. Journal of Machine Learning Research, 20(162):1–56.

[Arlot and Massart, 2009] Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. Journal of Machine learning research, 10(Feb):245–279.

[Bach, 2008] Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In Proceedings of the 25th international conference on Machine learning, pages 33–40.

[Baraud et al., 2009] Baraud, Y., , C., Huet, S., et al. (2009). Gaussian model selection with an unknown variance. The Annals of Statistics, 37(2):630–672.

[Barber and Candès, 2015] Barber, R. and Candès, E. (2015). Controlling the false discovery rate via knockoffs. The Annals of Statistics, 43(5):2055–2085.

[Barber et al., 2020] Barber, R., Candès, E., and Samworth, R. (2020). Robust inference with knockoffs. The Annals of Statistics, 48(3):1409–1431.

[Barber and Candès, 2019] Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. The Annals of Statistics, 47(5):2504–2537.

[Barski et al., 2007] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell, 129(4):823–837.

[Bartlett et al., 2012] Bartlett, P. L., Mendelson, S., and Neeman, J. (2012). ℓ_1-regularized linear regression: persistence and oracle inequalities. Probability theory and related fields, 154(1):193–224.

[Baudry et al., 2012] Baudry, J., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. Statistics and Computing, 22(2):455–470.

[Beck et al., 2005] Beck, V., Malick, J., and Peyré, G. (2005). Objectif agrégation, volume 8. H&K.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1):289–300.

[Benjamini and Hochberg, 2000] Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of educational and Behavioral Statistics, 25(1):60–83.

[Benjamini and Yekutieli, 2001] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, pages 1165–1188.

[Benjamini and Yekutieli, 2005] Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association, 100(469):71–81.

[Berk et al., 2013] Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. The Annals of Statistics, pages 802–837.

[Bertin et al., 2011] Bertin, K., Le Pennec, E., and Rivoirard, V. (2011). Adaptive Dantzig density estimation. Ann. Inst. Henri Poincaré Probab. Stat., 47(1):43–74.

[Bickel et al., 2009] Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2009). Simultaneous analysis of lasso and dantzig selector. The Annals of statistics, 37(4):1705–1732.

[Birgé and Massart, 2001] Birgé, L. and Massart, P. (2001). Gaussian model selection. Journal of the European Mathematical Society, 3(3):203–268.

[Birgé and Massart, 2007] Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. Probability theory and related fields, 138(1-2):33–73.

[Bogdan et al., 2013] Bogdan, M., Berg, E. v. d., Su, W., and Candes, E. (2013). Statistical estimation and testing via the sorted l1 norm. arXiv preprint arXiv:1310.1969.

[Bogdan et al., 2004] Bogdan, M., Ghosh, J. K., and Doerge, R. (2004). Modifying the schwarz bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics, 167(2):989–999.

[Bondell and Reich, 2012] Bondell, H. and Reich, B. (2012). Consistent high-dimensional bayesian variable selection via penalized credible regions. Journal of the American Statistical Association, 107(500):1610–1624.

[Bonferroni, 1936] Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8:3–62.

[Broman and Speed, 2002] Broman, K. W. and Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4):641–656.

[Bühlmann and Mandozzi, 2014] Bühlmann, P. and Mandozzi, J. (2014). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. Computational Statistics, 29(3):407–430.

[Bühlmann and Van De Geer, 2011] Bühlmann, P. and Van De Geer, S. (2011). Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media.

[Bunea et al., 2007a] Bunea, F., Tsybakov, A., and Wegkamp, M. (2007a). Sparsity oracle inequalities for the lasso. Electronic Journal of Statistics, 1:169–194.

[Bunea et al., 2006] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2006). Aggregation and sparsity via ℓ_1 penalized least squares. In International Conference on Computational Learning Theory, pages 379–391. Springer.

[Bunea et al., 2007b] Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2007b). Aggregation for gaussian regression. The Annals of Statistics, 35(4):1674–1697.

[Candes et al., 2007] Candes, E., Tao, T., et al. (2007). The dantzig selector: Statistical estimation when p is much larger than n. The annals of Statistics, 35(6):2313–2351.

[Candès et al., 2016] Candès, E. J., Fan, Y., Janson, L., and Lv, J. (2016). Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. Department of Statistics, Stanford University.

[Carré et al., 2017] Carré, C., Mas, A., and Krouk, G. (2017). Reverse engineering highlights potential principles of large gene regulatory network design and learning. NPJ systems biology and applications, 3(1):1–15.

[Castrillo et al., 2011] Castrillo, G., Turck, F., Leveugle, M., Lecharny, A., Carbonero, P., Coupland, G., Paz-Ares, J., and Oñate-Sánchez, L. (2011). Speeding cis-trans regulation discovery by phylogenomic analyses coupled with screenings of an arrayed library of arabidopsis transcription factors. PloS one, 6(6):e21524.

[Celeux et al., 2012] Celeux, G., El Anbari, M., Marin, J., and Robert, C. (2012). Regularization in regression: comparing bayesian and frequentist methods in a poorly informative situation. Bayesian Analysis, 7(2):477–502.

[Chen et al., 2018] Chen, D., Yan, W., Fu, L.-Y., and Kaufmann, K. (2018). Architecture of gene regulatory networks controlling flower development in arabidopsis thaliana. Nature communications, 9(1):1–13.

[Chen and Chen, 2008] Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. Biometrika, 95(3):759–771.

[Chen et al., 2006] Chen, X., Chen, M., and Ning, K. (2006). Bnarray: an r package for constructing gene regulatory networks from microarray data by using bayesian network. Bioinformatics, 22(23):2952–2954.

[Chen et al., 2021] Chen, Y., Jewell, S., and Witten, D. (2021). More powerful selective inference for the graph fused lasso.

[Chu and Corey, 2012] Chu, Y. and Corey, D. R. (2012). Rna sequencing: platform selection, experimental design, and data interpretation. Nucleic acid therapeutics, 22(4):271–274.

[Connault, 2011] Connault, P. (2011). Calibration d'algorithmes de type Lasso et analyse statistique de données métallurgiques en aéronautique. PhD thesis, Paris 11.

[Córdoba et al., 2019] Córdoba, I., Varando, G., Bielza, C., and Larrañaga, P. (2019). Generating random gaussian graphical models. arXiv preprint arXiv:1909.01062.

[Crowe et al., 2003] Crowe, M. L., Serizet, C., Thareau, V., Aubourg, S., Rouze, P., Hilson, P., Beynon, J., Weisbeek, P., Van Hummelen, P., Reymond, P., et al. (2003). Catma: a complete arabidopsis gst database. Nucleic acids research, 31(1):156–158.

[Cui et al., 2021] Cui, X., Dickhaus, T., Ding, Y., and Hsu, J. C. (2021). Handbook of multiple comparisons. CRC Press.

[De La Fuente et al., 2004] De La Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics, 20(18):3565–3574.

[Desboulets, 2018] Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. Econometrics, 6(4):45.

[Devijver, 2017] Devijver, E. (2017). Joint rank and variable selection for parsimonious estimation in a high-dimensional finite mixture regression model. Journal of Multivariate Analysis, 157:1–13.

[Devijver and Gallopin, 2018] Devijver, E. and Gallopin, M. (2018). Block-diagonal covariance selection for high-dimensional gaussian graphical models. Journal of the American Statistical Association, 113(521):306–314.

[Dudoit and van der Laan, 2008] Dudoit, S. and van der Laan, M. J. (2008). Multiple testing procedures with applications to genomics. Springer.

[Dunn, 1961] Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56(293):52–64.

[Durand, 2018] Durand, G. (2018). Multiple testing and post hoc bounds for heterogeneous data. Theses, Sorbonne Université.

[Duy and Takeuchi, 2021] Duy, V. and Takeuchi, I. (2021). More powerful conditional selective inference for generalized lasso by parametric programming. arXiv preprint arXiv:2105.04920.

[Efron, 2005] Efron, B. (2005). Local false discovery rates.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. The Annals of statistics, 32(2):407–499.

[Efron and Tibshirani, 1994] Efron, B. and Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

[Fan et al., 2014] Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. Annals of statistics, 42(1):324.

[Fan and Li, 2001] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456):1348–1360.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1.

[Gagnot et al., 2008] Gagnot, S., Tamby, J.-P., Martin-Magniette, M.-L., Bitton, F., Taconnat, L., Balzergue, S., Aubourg, S., Renou, J.-P., Lecharny, A., and Brunaud, V. (2008). Catdb: a public access to arabidopsis transcriptome data from the urgv-catma platform. Nucleic acids research, 36(suppl_1):D986–D990.

[Gégout-Petit et al., 2019] Gégout-Petit, A., Muller-Gueudin, A., and Karmann, C. (2019). The revisited knockoffs method for variable selection in l_1-penalised regressions.

[Geisser, 1975] Geisser, S. (1975). The predictive sample reuse method with applications. Journal of the American statistical Association, 70(350):320–328.

[Genovese and Wasserman, 2002] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3):499–517.

[Genovese et al., 2004] Genovese, C., Wasserman, L., et al. (2004). A stochastic process approach to false discovery control. The Annals of Statistics, 32(3):1035–1061.

[Genovese et al., 2006] Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. Biometrika, pages 509–524.

[Gibbs et al., 2003] Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The international hapmap project.

[Giraud, 2014] Giraud, C. (2014). Introduction to high-dimensional statistics, volume 138. CRC Press.

[Giraud et al., 2012] Giraud, C., Huet, S., Verzelen, N., et al. (2012). High-dimensional regression with unknown variance. Statistical Science, 27(4):500–518.

[Goeman and Solari, 2014] Goeman, J. J. and Solari, A. (2014). Multiple hypothesis testing in genomics. Statistics in medicine, 33(11):1946–1978.

[Goldfarb and Idnani, 1983] Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. Mathematical programming, 27(1):1–33.

[Golub et al., 1999] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439):531–537.

[Guo and Rao, 2008] Guo, W. and Rao, M. B. (2008). On control of the false discovery rate under no assumption of dependency. Journal of Statistical Planning and Inference, 138(10):3176–3188.

[Haury et al., 2012] Haury, A., Mordelet, F., Vera-Licona, P., and Vert, J. (2012). Tigress: trustful inference of gene regulation using stability selection. BMC systems biology, 6(1):145.

[Hilson et al., 2004] Hilson, P., Allemeersch, J., Altmann, T., Aubourg, S., Avon, A., Beynon, J., Bhalerao, R. P., Bitton, F., Caboche, M., Cannoot, B., et al. (2004). Versatile gene-specific sequence tags for arabidopsis functional genomics: transcript profiling and reverse genetics applications. Genome research, 14(10b):2176–2189.

[Hochberg, 1988] Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. Biometrika, 75(4):800–802.

[Hoerl and Kennard, 1988] Hoerl, A. and Kennard, R. (1988). Ridge regression, in 'encyclopedia of statistical sciences', vol. 8.

[Holm, 1979] Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, pages 65–70.

[Hommel, 1988] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. Biometrika, 75(2):383–386.

[Huang, 2008] Huang, C. (2008). Risk of penalized least squares, greedy selection and ell1-penalization for flexible function libraries. Yale University.

[Huang and Janson, 2020] Huang, D. and Janson, L. (2020). Relaxing the assumptions of knockoffs by conditioning. The Annals of Statistics, 48(5):3021–3042.

[Hyun et al., 2018] Hyun, S., G'sell, M., and Tibshirani, R. (2018). Exact post-selection inference for the generalized lasso path. Electronic Journal of Statistics, 12(1):1053–1097.

[Ismaili and Gaillard, 2009] Ismaili, A. and Gaillard, P. (2009). Le lasso, ou comment choisir parmi un grand nombre de variables à l'aide de peu d'observations. Exposé de maîtrise.

[Ivanoff et al., 2016] Ivanoff, S., Picard, F., and Rivoirard, V. (2016). Adaptive Lasso and group-Lasso for functional Poisson regression. J. Mach. Learn. Res., 17:Paper No. 55, 46.

[Jacob et al., 2009] Jacob, L., Obozinski, G., and Vert, J. (2009). Group lasso with overlap and graph lasso. In Proceedings of the 26th annual international conference on machine learning, pages 433–440.

[Kos and Bogdan, 2020] Kos, M. and Bogdan, M. (2020). On the asymptotic properties of slope. Sankhya A, 82(2):499–532.

[Kschischang, 2017] Kschischang, F. R. (2017). The complementary error function. Online, April.

[Laurent and Massart, 2000] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. Annals of Statistics, pages 1302–1338.

[Lauritzen, 1996] Lauritzen, S. (1996). Graphical Models. Oxford University Press.

[Lebarbier, 2005] Lebarbier, E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. Signal processing, 85(4):717–736.

[Lee et al., 2016] Lee, J., Sun, D., Sun, Y., and Taylor, J. (2016). Exact post-selection inference, with application to the lasso. The Annals of Statistics, 44(3):907–927.

[Lehmann and Romano, 2005] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. Ann. Statist., 33(3):1138–1154.

[Lenoir and Giannella, 2006] Lenoir, T. and Giannella, E. (2006). The emergence and diffusion of dna microarray technology. Journal of biomedical discovery and collaboration, 1(1):1–39.

[Lerasle, 2009] Lerasle, M. (2009). Rééchantillonnage et sélection de modèles optimale pour l'estimation de la densité. PhD thesis, Toulouse, INSA.

[Leung and Sun, 2021] Leung, D. and Sun, W. (2021). Zap: z-value adaptive procedures for false discovery rate control with side information. arXiv preprint arXiv:2108.12623.

[Lim and Yu, 2016] Lim, C. and Yu, B. (2016). Estimation stability with cross-validation (escv). Journal of Computational and Graphical Statistics, 25(2):464–492.

[Liu et al., 2012] Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric gaussian copula graphical models. The Annals of Statistics, 40(4):2293–2326.

[MacArthur et al., 2017] MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). Nucleic acids research, 45(D1):D896–D901.

[Maertens et al., 2018] Maertens, A., Tran, V., Kleensang, A., and Hartung, T. (2018.). Weighted gene correlation network analysis (wgcna) reveals novel transcription factors associated with bisphenol a dose-response. Frontiers in Genetics., 9(208).

[Mallows, 2000] Mallows, C. L. (2000). Some comments on cp. Technometrics, 42(1):87–94.

[Marchini et al., 2005] Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature genetics, 37(4):413–417.

[Massart and Meynet, 2010] Massart, P. and Meynet, C. (2010). An l1-oracle inequality for the lasso. arXiv preprint arXiv:1007.4791.

[Meinshausen, 2007] Meinshausen, N. (2007). Relaxed lasso. Computational Statistics & Data Analysis, 52(1):374–393.

[Meinshausen and Bühlmann, 2010] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4):417–473.

[Meinshausen et al., 2006] Meinshausen, N., Bühlmann, P., et al. (2006). High-dimensional graphs and variable selection with the lasso. The annals of statistics, 34(3):1436–1462.

[Meinshausen et al., 2009] Meinshausen, N., Yu, B., et al. (2009). Lasso-type recovery of sparse representations for high-dimensional data. The annals of statistics, 37(1):246–270.

[Meynet and Maugis-Rabusseau, 2012] Meynet, C. and Maugis-Rabusseau, C. (2012). A sparse variable selection procedure in model-based clustering.

[Mitsuda and Ohme-Takagi, 2009] Mitsuda, N. and Ohme-Takagi, M. (2009). Functional analysis of transcription factors in arabidopsis. Plant and Cell Physiology, 50(7):1232–1248.

[Morin et al., 2008] Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J., and Marra, M. A. (2008). Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. Biotechniques, 45(1):81–94.

[Natarajan, 1995] Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. SIAM journal on computing, 24(2):227–234.

[O'Malley et al., 2016] O'Malley, R. C., Huang, S.-s. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A., and Ecker, J. R. (2016). Cistrome and epicistrome features shape the regulatory dna landscape. Cell, 165(5):1280–1292.

[Peng and Wang, 2015] Peng, B. and Wang, L. (2015). An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. Journal of Computational and Graphical Statistics, 24(3):676–694.

[Peng et al., 2009] Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association, 104(486):735–746.

[Pötscher, 1991] Pötscher, B. M. (1991). Effects of model selection on inference. Econometric Theory, 7(2):163–185.

[Razaghi-Moghadam and Nikoloski, 2020] Razaghi-Moghadam, Z. and Nikoloski, Z. (2020). Supervised learning of gene-regulatory networks based on graph distance profiles of transcriptomics data. NPJ systems biology and applications, 6(1):1–8.

[Refaeilzadeh et al., 2009] Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-validation. Encyclopedia of database systems, 5:532–538.

[Rigollet and Tsybakov, 2011] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. The Annals of Statistics, 39(2):731–771.

[Rom, 1990] Rom, D. M. (1990). A sequentially rejective test procedure based on a modified bonferroni inequality. Biometrika, 77(3):663–665.

[Romano et al., 2008] Romano, J. P., Shaikh, A. M., and Wolf, M. (2008). Control of the false discovery rate under dependence using the bootstrap and subsampling. Test, 17(3):417–442.

[Romano et al., 2007] Romano, J. P., Wolf, M., et al. (2007). Control of generalized error rates in multiple testing. The Annals of Statistics, 35(4):1378–1408.

[Roquain, 2011] Roquain, E. (2011). Type i error rate control for testing many hypotheses: a survey with proofs. Journal de la Société Française de Statistique, 152(2):3–38.

[Saporta, 2006] Saporta, G. (2006). Probabilités, analyse des données et statistique. Editions technip.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. The annals of statistics, 6(2):461–464.

[Segal et al., 2003] Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003). Regression approaches for microarray data analysis. Journal of Computational Biology, 10(6):961–980.

[Simes, 1986] Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. Biometrika, 73(3):751–754.

[Stone, 1974] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological), 36(2):111–133.

[Storey et al., 2004] Storey, J., Taylor, J., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(1):187–205.

[Storey, 2002] Storey, J. D. (2002). A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3):479–498.

[Storey et al., 2003] Storey, J. D. et al. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. The Annals of Statistics, 31(6):2013–2035.

[Su et al., 2017] Su, W., Bogdan, M., Candes, E., et al. (2017). False discoveries occur early on the lasso path. The Annals of statistics, 45(5):2133–2150.

[Swarbreck et al., 2007] Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., et al. (2007). The arabidopsis information resource (tair): gene structure and function annotation. Nucleic acids research, 36(suppl_1):D1009–D1014.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

[Tibshirani et al., 2005] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108.

[Tibshirani et al., 2016] Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association, 111(514):600–620.

[Van de Geer et al., 2011] Van de Geer, S., Bühlmann, P., and Zhou, S. (2011). The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). Electronic Journal of Statistics, 5:688–749.

[Vasseur, 2017] Vasseur, Y. (2017). Inférence de réseaux de régulation orientés pour les facteurs de transcription d'Arabidopsis thaliana et création de groupes de co-régulation. PhD thesis, Université Paris-Saclay (ComUE).

[Verzelen et al., 2012] Verzelen, N. et al. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. Electronic Journal of Statistics, 6:38–90.

[Vinga, 2021] Vinga, S. (2021). Structured sparsity regularization for analyzing high-dimensional omics data. Briefings in Bioinformatics, 22(1):77–87.

[Wainwright, 2009] Wainwright, M. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using l_1-constrained quadratic programming. IEEE Transactions on Information Theory, 55(5):2183–2202.

[Wang et al., 2020] Wang, F., Mukherjee, S., Richardson, S., and Hill, S. (2020). High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking. Statistics and computing, 30(3):697–719.

[Wang et al., 2012] Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. Journal of the American Statistical Association, 107(497):214–222.

[Wu and Ma, 2015] Wu, C. and Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics. Briefings in bioinformatics, 16(5):873–883.

[Wu and Ye, 2006] Wu, X. and Ye, Y. (2006). Exploring gene causal interactions using an enhanced constraint-based method. Pattern Recognition, 39(12):2439–2449.

[Yang et al., 2017] Yang, H., Li, S., Cao, H., Zhang, C., and Cui, Y. (2017). Predicting disease trait with genomic data: a composite kernel approach. Briefings in Bioinformatics, 18(4):591–601.

[Yang, 2005] Yang, Y. (2005). Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. Biometrika, 92(4):937–950.

[Yuan and Lin, 2006] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1):49–67.

[Zaag et al., 2015] Zaag, R., Tamby, J. P., Guichard, C., Tariq, Z., Rigaill, G., Delannoy, E., Renou, J.-P., Balzergue, S., Mary-Huard, T., Aubourg, S., et al. (2015). Gem2net: from gene expression modeling to-omics networks, a new catdb module to investigate arabidopsis thaliana genes involved in stress response. Nucleic acids research, 43(D1):D1010–D1017.

[Zhang et al., 2022] Zhang, D., Khalili, A., and Asgharian, M. (2022). Post-model-selection inference in linear regression models: An integrated review. Statistics Surveys, 16:86–136.

[Zhao et al., 2009] Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. The Annals of Statistics, 37(6A):3468–3497.

[Zhao and Yu, 2006] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. Journal of Machine learning research, 7(Nov):2541–2563.

[Zhou, 2009] Zhou, S. (2009). Thresholding procedures for high dimensional variable selection and statistical estimation. Advances in Neural Information Processing Systems, 22:2304–2312.

[Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476):1418–1429.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2):301–320.

[Zou et al., 2007] Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. The Annals of Statistics, 35(5):2173–2192.

[Zou and Yuan, 2008] Zou, H. and Yuan, M. (2008). Regularized simultaneous model selection in multiple quantiles regression. Computational Statistics & Data Analysis, 52(12):5296–5304.

List of Figures

List of Tables

List of Algorithms