Contrôle des fausses découvertes lors de la sélection de variables en grande dimension

Jean-Michel Bécu

▶ **To cite this version:**

Jean-Michel Bécu. Contrôle des fausses découvertes lors de la sélection de variables en grande dimension. Autre [cs.OH]. Université de Technologie de Compiègne, 2016. Français. NNT : 2016COMP2264 . tel-01326950v2

HAL Id: tel-01326950

https://theses.hal.science/tel-01326950v2

Submitted on 24 Oct 2016

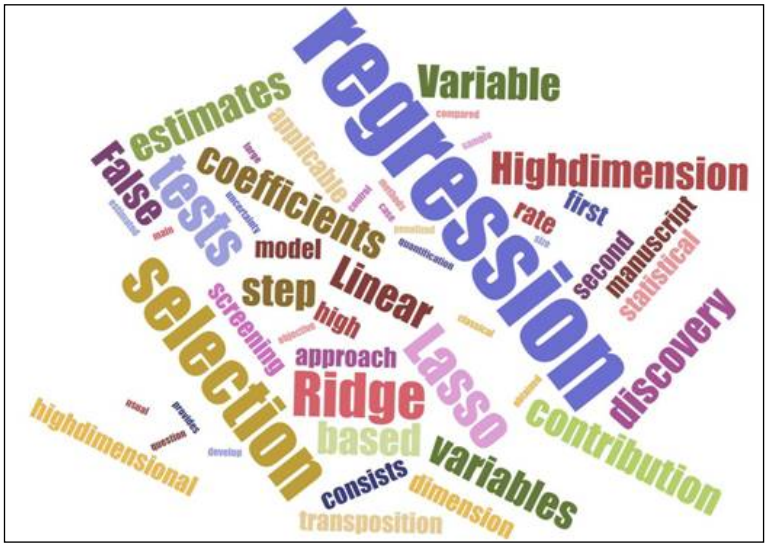
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Par **Jean-Michel BÉCU**

Contrôle des fausses découvertes lors de la sélection de variables en grande dimension

Thèse présentée
pour l'obtention du grade
de Docteur de l'UTC



Soutenue le 10 mars 2016
Spécialité : Technologies de l'Information et des Systèmes

D2264

Université de Technologie de
Compiègne
Heudiasyc

Thèse

En vue de l'obtention du titre de Docteur
Spécialité « Technologies de l'Information et des Systèmes »

présentée par

Jean-Michel Bécu

Contrôle des fausses découvertes lors de la sélection de variables en grande dimension

Thèse soutenue le 10 mars 2016 devant le jury composé de :

Rapporteur : Antoine CORNUÉJOLS, AgroParisTech, Professeur des Universités
Rapporteur : Alain RAKATOMAMONJY, Université de Rouen, Professeur des Universités
Examinateur : Gérard GOVAERT, Université de Technologie de Compiègne, Professeur des Universités
Examinateur : Éric LECLERC, Université de Tokyo, Directeur de recherche CNRS
Examinateur : Stéphane ROBIN, AgroParisTech, Directeur de Recherche INRA
Examinateur : Étienne ROQUAIN, Université Pierre et Marie Curie, Maître de conférence
Directeur : Christophe AMBROISE, Université d'Évry Val d'Essonne, Professeur des Universités
Directeur : Yves GRANDVALET, Université de Technologie de Compiègne, Directeur de recherche

Control of false discoveries in high-dimensional variable selection

A Maria,

"Scientists have calculated that the chances of something so patently absurd actually existing are millions to one. But magicians have calculated that million-to-one chances crop up nine times out of ten." Terry Pratchett

Titre Contrôle des fausses découvertes lors de la sélection de variables en grande dimension

Résumé Dans le cadre de la régression, de nombreuses études s'intéressent au problème dit de la grande dimension, où le nombre de variables explicatives mesurées sur chaque échantillon est beaucoup plus grand que le nombre d'échantillons. Si la sélection de variables est une question classique, les méthodes usuelles ne s'appliquent pas dans le cadre de la grande dimension. Ainsi, dans ce manuscrit, nous présentons la transposition de tests statistiques classiques à la grande dimension. Ces tests sont construits sur des estimateurs des coefficients de régression produits par des approches de régressions linéaires pénalisées, applicables dans le cadre de la grande dimension. L'objectif principal des tests que nous proposons consiste à contrôler le taux de fausses fausses découvertes. La première contribution de ce manuscrit répond à un problème de quantification de l'incertitude sur les coefficients de régression réalisée sur la base de la régression ridge, qui pénalise les coefficients de régression par leur norme $l2$, dans le cadre de la grande dimension. Nous y proposons un test statistique basé sur le rééchantillonage. La seconde contribution porte sur une approche de sélection en deux étapes : une première étape de criblage des variables, basée sur la régression parcimonieuse Lasso précède l'étape de sélection proprement dite, où la pertinence des variables pré-sélectionnées est testée. Les tests sont construits sur l'estimateur de la régression ridge adaptive, dont la pénalité est construite à partir des coefficients de régression du Lasso. Une dernière contribution consiste à transposer cette approche à la sélection de groupes de variables.

Mots-clés Sélection de variables, Grande dimension, Taux de fausses découvertes, Modèle linéaire, Ridge, Lasso, Méthodes en deux étapes

Title Control of False Discoveries in High-Dimensional Variable Selection

Abstract In the regression framework, many studies are focused on the high-dimensional problem where the number of measured explanatory variables is very large compared to the sample size. If variable selection is a classical question, usual methods are not applicable in the high-dimensional case. So, in this manuscript, we develop the transposition of statistical tests to the high dimension. These tests operate on estimates of regression coefficients obtained by penalized linear regression, which is applicable in high-dimension. The main objective of these tests is the false discovery control. The first contribution of this manuscript provides a quantification of the uncertainty for regression coefficients estimated by ridge regression in high dimension. The ridge regression penalizes the coefficients on their $l2$ norm. To do this, we devise a statistical test based on permutations. The second contribution is based on a two-step selection approach. A first step is dedicated to the screening of variables, based on parsimonious regression Lasso. The second step consists in cleaning the resulting set by testing the relevance of pre-selected variables. These tests are made on adaptive-ridge estimates, where the penalty is constructed on lasso estimates learned during the screening step. A last contribution consists to the transposition of this approach to group-variables selection.

Keywords Variable selection, High-dimension, False discovery rate, Linear model, Ridge, Lasso, Two-step approaches

Table des matières

Liste des tableaux

Table des figures

Remerciements

Si, il est assez facile de résumer trois ans de thèse d'un point de vue scientifique (quoique), il est très difficile de le faire d'un point de vue humain et affectif. L'humeur d'un doctorant est comme le cours de la bourse, il y a des embellies, des remontées mais surtout il y a des krachs. J'avais toujours la chanson de robin de bois dans la tête : "Des bas des hauts, il y en a partout, mais des drames, il y en a surtout, ...". Une thèse, c'est l'apprentissage de l'éternelle insatisfaction couplée à une forme de jubilation. Avant de regarder les résultats de chaque simulation, j'étais tel un enfant le jour de noël qui va ouvrir son cadeau, prêt a sauter dans tous les sens de joie, de disséminer l'ordinateur façon puzzle ou au pire d'utiliser le cordon d'alimentation version nœud coulant. Heureusement les moments de joie ont été les plus nombreux. Une thèse c'est avant tout une aventure, une sorte de livre dont vous êtes le héros. Si ce travail est en mon nom il n'aurait pas pu être possible sans ceux qui m'ont accompagné et que je me dois de remercier. Et je commencerai tout d'abord par mes encadrant car une thèse ne peut être réussie que si l'environnement de travail est propice et grâce à eux il l'a été.

Je tiens en premier lieu à remercier Yves, mon directeur de thèse. Je lui en ai fait voir pendant ces trois ans et malgré ça il a toujours été présent. Une fois par semaine il venait nous voir à Évry et nous étions partis pour une demi-journée voir une journée de réunion, entrecoupée par la dégustation d'un fondant au chocolat accompagné d'une petite glace (pour se recharger les batteries). Je me rappelle qu'au bout de 6 mois de thèse Yves s'était demandé si j'étais un génie qui méritait la médaille Field, car j'avais réussi à expliquer le bruit. Mais au fi-

nal c'était plutôt l'idée d'un abruti. Enfin abruti pas tant que ça, car nous avons produit de belles choses et sans un guide comme Yves cela n'aurait pas été possible. De plus il faut noter qu'il est un des rares à avoir un décodeur de Jean-Michel intégré, et donc de me comprendre quand je partait dans des explications floues. Ou du moins il me demandait de recommencer en considérant que je parlais à quelqu'un et non pas à moi-même. Il y aurait beaucoup d'autres choses à dire pour le remercier à sa juste valeur, mais cela restera entre nous.

Je tiens en "second" premier lieu à remercier Christophe, mon autre directeur de thèse. Je l'ai contacté en 2009 pour pouvoir faire mon apprentissage à stat & génome (heu non au LaMME ...) mais cela n'avait pas été possible pour des raisons financières. Comme dit la chanson "On s'était donné rendez-vous dans 2 ans" et à la fin de mon master je l'ai recontacté. Imaginez-la scène : Moi "Bonjour, je suis un biologiste qui à fait un master de bioinfo et je voudrais faire une thèse de math" et lui "Ok on tente le coup". Il s'est démené pour me trouver un financement, et en plus du financement il m'a trouvé un Yves. Que demande le peuple ? Merci, de m'avoir fait confiance, donné ma chance et soutenu durant ces trois ans. Merci pour avoir pris le temps de m'accompagner dans mon apprentissage du sujet et pour tous les conseils pour mener à bien cette activité de recherche. C'était une joie et un honneur qu'il soit là le jour de mon mariage et pour la trace qu'il m'y a offert. Encore ici, il y aurait beaucoup d'autres choses à dire pour le remercier à sa juste valeur, mais cela restera entre nous.

En premier second lieu je remercie l'ensemble des membres du jury d'avoir accepté de juger l'ensemble de mon travail et l'avoir sanctionné du titre de docteur. Plus particulièrement Antoine Cornuéjols et Alain Rakatomamonjy d'avoir été mes rapporteurs, Gérard Govaert de l'avoir présidé et Stéphane Robin et Étienne Roquain d'en avoir été examinateur avec autant de questions que des rapporteurs. Vos nombreuses questions et la discussion sur mon travail ont été extrêmement intéressantes et permettent d'ouvrir de nouvelles pistes pour des travaux futurs. Merci également à Éric Leclerc qui n'a pu être présent le jour

de la soutenance mais qui a tout de même partagé son avis sur mon travail.

Avant de parler de l'ensemble de mes collègues, je tiens à remercier Dominique Cellier qui a été mon professeur de biostatistique à l'université de Rouen. Il m'a suivi tout au long de mon cursus et c'est à cause (grâce) à lui que j'ai souhaité bifurquer vers cette spécialité et que j'ai pu rentrer en contact avec l'équipe stat & génome. Il est de ces personnes qui sont capables de changer votre vie. Pour parler de lui, un seul mot me vient c'est celui de "mentor". Alors je le remercie pour tout cela et pour son amitié.

Pour éviter un oubli, je remercie l'ensemble des collègues du LaMME que j'ai côtoyé pendant toute ma thèse. Grâce à vous tous j'ai passé trois superbes années autant pour les échanges scientifiques que pour ceux qui l'étaient moins. L'ordre des remerciements suivant n'est en aucun cas un classement hiérarchique de vos importances.

Un grand merci à Cyril pour avoir participer de manière active dans mon travail de recherche, cela nous a permis de nous associer pour un article. Je tiens à mettre en avant son coté humain, sa simplicité et sa curiosité sincère pour les autres.

Merci également à Pierre pour tous les échanges que nous avons eu, et pour avoir su et pris le temps de répondre à mes nombreuses questions même si elles étaient stupides. Ce sera un plaisir si nos chemins se recroisent dans le travail. J'en profite pour remercier une nouvelle fois Étienne qui m'a également énormément apporté pour les problèmes statistiques.

Merci également à Carène, qui apporte ce coté bioinfo au laboratoire. Pour sa gentillesse et les attentions qu'elle a pu avoir durant toute ma thèse.

Ne pouvant remercier tout les doctorants du laboratoire car ce serait trop long, je remercie tous les compagnons de galère. Mais plus spécifiquement, je remercie Morgane qui a été ma première et ma dernière collègue de bureau. Cela a été un honneur de partager ces moments de partage scientifiques mais surtout un plaisir d'avoir pu partager son

amitié dans et en dehors du laboratoire (vive les soirées jeux). Ce fut également un plaisir de la voir devenir mère quand moi je devenais père pour la deuxième fois et Pierre pour la troisième. Vivre des grossesses en parallèle cela soude encore plus les gens, par le partage d'une expérience si forte au même moment. Pour être moins formel : Courage Morgane, la fin est proche, des personnes ont la recherche dans le sang alors vas-y coloc !

Un merci aussi à Sarah qui a soutenu sa thèse il y a un an et que j'ai torturé de questions, de coup de blues et surtout que je prenais plaisir à faire sursauter.

Je tiens à remercier Michèle notre secrétaire, pour nos pauses cigarettes et pour avoir toujours su s'arranger avec mon allergie administrative. Et aussi pour ces tranches de vies partagées où l'on se rend compte que le monde est bien petit.

Un grand merci à Maurice, notre administrateur système, qui m'a permis de continuer à Geeker un peu, qui m'a fait saliver vidéo-ludiquement parlant et pour tout nos fous rires.

Un grand et énorme merci à Jean-Pierre, qui se reconnaitra. Il va me manquer mais on se reverra.

J'allais oublier de remercier tous les participants du badminton du mercredi pour ces moments de rigolades et de victoires. Ainsi je remercie Morgane, Pierre, Cyril, le fantôme Maurice, Michael mais je ne remercie pas Étienne car lui ne me laissait pas gagner.

Pour terminer je remercierai pudiquement ma famille. Mes parents et mon frère pour m'avoir soutenu durant la thèse, d'avoir accepté de la relire et d'avoir été là pour la soutenance. Il était important que vous soyez-là pour je cite "avoir la reconnaissance de mes pairs". Un merci aussi à Gilles et Cathy d'avoir fait la route depuis Martignat.

Mais après tout cela il reste le plus important : ma femme Magali qui a été mon support et mes deux filles Maria et Charlotte qui ont été mes moteurs. Tout ce petit monde a vécu/subi/permis cette aventure. Nos filles en sont le symbole avec Maria née un mois après le début ma thèse et Charlotte née pendant la rédaction. Merci à vous pour la

ressource inépuisable de joie que vous m'apportez et pour le chemin qu'il nous reste à parcourir qui statistiquement a 100% de chance d'être magnifique.

Abréviations & Notations

Abréviations

Ci-dessous, la liste des abréviations utilisées dans ce manuscrit, le terme en anglais plus commun est entre parenthèse si il existe :

ADN : acide désoxyribonucléique *(deoxyribonucleic acid)*

AR : moindres carrés pénalisés adaptatifs *(adaptive-ridge)*

BH : Procédure de Benjamini-Hochberg

BLOCK : dessin corrélé

CRL : Cluster Representative Lasso

CV : Validation croisée *(Cross-Validation)*

df : degrés de liberté *(degrees of freedom)*

FDP : Proportion de fausses découvertes *(False Discovery Proportion)*

FDR : taux de fausses découvertes *(False Discovery Rate)*

FP : Faux Positif *(False Positive - FP)*

FPR : taux de faux positifs ou erreur de type I *(False Positive Rate)*

FN : Faux Négatif *(False Negative - FN)*

FNR : taux de faux négatif ou erreur de type II *(False Negative Rate)*

FWER : taux d'erreur par famille *(Family-Wise Error Rate)*

GL : group-Lasso

GROUP : dessin corrélé, variables pertinentes regroupées

GWAS : étude d'association pangénomique *(Genome-Wise Association Study)*

IC/CI : intervalle de confiance *(Confidence Interval)*

IND : dessin non corrélé

Lasso : régression pénalisée en norme $l1$ *(Least Absolute Shrinkage and Selection Operator)*

MHC : Complexe majeur d'histocompatibilité *(Major Histocompatibility Complex)*

MSE : erreur quadratique moyenne *(Mean Squared Error)*

OLS : régression des moindres carrés *(Ordinary Least Squares)*

RSS : Somme des carrés des résidus *(Residual Sum of Square)*

SEN : Sensibilité *(Sensitivity)*

SGL : sparse-group-Lasso

SNP : polymorphisme nucléotidique *(Single-Nucleotide Polymorphisms)*

SNR : rapport signal sur bruit *(Signal-to-Noise Ratio)*

TOEP$^-$: dessin corrélé négativement selon une matrice de Toeplitz, variables pertinentes regroupées

VP : Vrai Positif *(True Positive - TP)*

VN : Vrai Négatif *(True Negative - TN)*

Notations

Ci-dessous, la liste des notations couramment utilisées dans ce manuscrit :

$\mathcal{H}_0$: hypothèse nulle

$\mathcal{H}_1$: hypothèse alternative

n : nombre d'observations, taille de l'échantillon

p : nombre de variables explicatives

G : nombre de groupes de variables

x : variable aléatoire

$\mathbf{x}$: vecteur $(x_1, \cdots, x_p)$

$\mathbf{X}$: matrice $(\mathbf{x}_1, \cdots, \mathbf{x}_p)$

$|\mathbf{x}|$: taille du vecteur $\mathbf{x}$

$\|\mathbf{x}\|_l$: norme l du vecteur $\mathbf{x}$

$\mathbf{y}$: vecteur réponse

$\hat{\mathbf{y}}$: estimation du vecteur réponse

$\boldsymbol{\epsilon}$: bruit gaussien

$\mathcal{D}$: jeux de données correspondant à $(\mathbf{x}_i, y_i)$ pour $i \in \{1, ..., n\}$

α : risque d'erreur

P : p-valeur

$\boldsymbol{\Omega}$: modèle complet comprenant l'ensemble des variables

$\boldsymbol{\omega}$: sous-modèle de $\boldsymbol{\Omega}$, appelé modèle réduit

$\bar{\boldsymbol{\omega}}$: complémentaire du sous-modèle $\boldsymbol{\omega}$

$\boldsymbol{\beta}^\star$: paramètres/coefficients des variables

$\hat{\boldsymbol{\beta}}$: estimation des coefficients $\boldsymbol{\beta}^*$

λ_k : coefficient de penalité pondérant la norme k des coefficients

$\mathcal{S}^\star$: support : ensemble des variables $j : \{\boldsymbol{\beta}_j \neq 0\}$

$\hat{\mathcal{S}}$: ensemble des variables sélectionnées durant l'étape de criblage (*screening*)

$\hat{\mathcal{S}}(\hat{\lambda})$: ensemble des variables pour lesquelles $\mathcal{H}_0$ est rejetée durant l'étape de nettoyage (*cleaning*)

r^2 : coefficient de corrélation partielle

t : statistique de Student

R : résidus

B : Nombre de réplications dans une procédure de rééchantillonage

$\tilde{\mathbf{X}}$: matrice dite de knockoff

O : complexité

cov : covariance

var : variance

Introduction

Cette thèse s'inscrit dans l'étude des problématiques inhérentes au phénomène de la grande dimension. Nous y présenterons des approches méthodologiques et statistiques permettant d'y faire de la sélection ainsi que les apports et innovation que nous avons apportés au domaine.

Au début de cette thèse, financée dans le cadre du projet ECOTOX de la fondation UTC, nous souhaitions analyser des données de mesure d'expression de gènes sur cellules cultivées à l'aide de la technologie des biopuces (Snouber et al. 2012). Cette technologie, basée sur des microcanaux, permet d'induire des cinétiques de fluides sur le milieu de culture et ainsi de reproduire de manière plus réaliste l'environnement auquel sont confrontées ces cellules *in vivo*. Ce procédé *in vitro* est totalement innovant car à l'heure actuelle les cultures cellulaires ont essentiellement lieu sur des boîtes de Pétri, où la dynamique de l'environnement cellulaire est totalement absente.

Plus précisément, notre objectif était d'analyser les données d'expressions de gènes, obtenues par des analyses *micro-array*, pour en extraire les gènes différentiellement exprimés, entre deux conditions expérimentales, par exemple entre les biopuces et les boîtes de pétri afin de voir l'apport de cette technologie. Des analyses sur les effets de toxine ou de médicaments sur des lignées cellulaires spécifiques devaient également être menées. Les puces utilisées pour les analyses recensaient au moins 20.000 gènes, et après des pré-traitements "classiques" Hochreiter et al. (2006), il restait au moins 5.000 gènes pour chaque analyse alors que le nombre d'échantillons se situait entre 6 et 12 en fonction des analyses. Extraire de l'information pertinente de

ces données sous-entendait donc de gérer le problème de la grande dimension où le nombre de variables (les gènes) dépassait largement le nombre d'échantillons (les expériences). Les solutions qui seront présentées dans ce manuscrit n'ont finalement pas permis d'extraire des informations pertinentes pour les données de biopuces. En effet, nos protocoles s'y appliquent mais le trop faible nombre d'échantillons n'a pas permis de retourner des résultats à des niveaux de confiances acceptables.

Une approche de sélection standard pour ce type de données, consiste à tester de manière indépendante, l'expression de chaque gène sur les différents échantillons, et ensuite de regrouper tous ces tests en prenant en compte leur multiplicité pour la sélection globale. Ce genre d'approche risque de ne pas être très sensible et surtout on y fait l'hypothèse que les effets des gènes sont indépendants. Pourtant, en réalité les modèles biologiques font l'hypothèse d'une accumulation/coordination d'expression de gènes plus ou moins fortement co-régulés. C'est ainsi que, considérant un modèle statistique multivarié avec un nombre d'échantillons faible devant le nombre de variables, nous nous sommes dirigés vers les méthodes de régression pénalisée et plus spécifiquement la régression Lasso (Tibshirani 1996). L'estimation de la pertinence des gènes se fait alors de manière globale, afin d'estimer les paramètres du modèle.

Ces approches sont de bonnes candidates pour ce type d'analyse et depuis quelques années la communauté scientifique s'intéresse à trouver des solutions pour mesurer la significativité des estimateurs en régression Lasso (Wasserman and Roeder 2009). Toutefois, au début de la thèse, les tests statistiques construits sur les estimateurs du Lasso n'étais pas des tests exacts, c'est à dire calculant exactement une probabilité selon la loi que suivent ces variables aléatoires. Il est a noté que les tests statistiques sur les estimateurs du Lasso sont le sujet de nombreuses études, et que durant ces trois ans un certains nombres de solutions on été proposées (Lockhart et al. 2014, Barber and Candes 2014) mais elles ne s'appliquent pas en grande dimension. D'un autre

côté, il apparaîssait dans la bibliographie que la régression ridge, une autre méthode de régression pénalisée, offrait plus facilement la capacité d'y effectuer des tests statistiques (Halawa and El Bassiouni 1999, Cule et al. 2011).

L'idée de base de la thèse fut d'associer les propriétés de sélection du Lasso et les tests statistiques sur la ridge par des liens existants entre ces deux méthodes de régression (Grandvalet 1998). Toutefois, les tests "classiques" que nous utilisions dans le cadre de la régression ridge ne sont pas exacts et leurs approximations ne sont pas valides en grande dimension. A ce stade nous avions déjà mis en place un protocole complet en deux étapes, construit autour du lien entre le Lasso et la ridge et permettant la mise en place de procédure de test sur les variables candidates. Nous avons alors recherché des solutions alternatives permettant de tester les estimateurs de la régression ridge.

Ce manuscrit de thèse s'organise donc autour de ces différents éléments, où deux problématiques apparaîtront comme un fil rouge. La première problématique sera la notion de test statistique sur les estimateurs obtenus et plus précisément le contrôle du taux de faux positifs sur les variables sélectionnées, c'est à dire le taux de fausses découvertes. La seconde problématique sera bien évidemment la problématique de la grande dimension qui complique la tâche précédente. Ainsi, ce manuscrit sera composé de quatre chapitres organisés comme suit :

État de l'art Dans ce chapitre seront présentés tout les outils utilisés durant le manuscrit et les connaissances/limites existantes dans le domaine de la sélection de variables en grande dimensions. Cela commencera par une présentation des théories et méthodologies existantes sur l'intérêt de la sélection de variable en biologie ainsi que des questions statistiques associées. Ceci correspondra à la première problématique où sera mis en exergue la question de la définition d'une variable effectivement explicative. Cela sera suivi par une présentation des théories et méthodologies existantes sur les méthodes de régressions allant de celles applicables à la petite dimension jusqu'à celles en grande

dimension. Ceci correspond à la seconde problématique ainsi qu'aux effets de la grande dimension expliquant la difficulté d'y construire des tests statistiques.

Régression ridge et tests statistiques Ce chapitre expliquera la difficulté de construire des tests exacts sur la régression ridge, et dressera un panorama des méthodes existantes ou des approches possibles permettant d'effectuer ces tests. Il y sera également proposé une méthode de permutation permettant d'effectuer des tests approchés utilisables en grande dimension sur les estimateurs de la régression ridge sans qu'il y ait de contrainte particulière sur la force de la pénalité utilisée. Les travaux correspondant à ce chapitre n'ont pas encore été intégralement soumis à publication.

Sélection de variable par l'*adaptive-ridge* Ce chapitre présentera une approche en deux étapes inspirée de la méthode "screen and clean" de Wasserman and Roeder (2009), permettant de sélectionner un sous ensemble de variables d'intérêt en y contrôlant le taux d'erreur. Cette procédure, basé sur l'idée de sous-échantillonnage des données (Cox 1975), effectue une pré-sélection de variables candidates par le Lasso suivie par un test sur ces variables. Les améliorations que nous avons apportées et surtout notre utilisation de l'*adaptive-ridge* y seront discutés. Ces travaux ont été publiés (Becu et al. 2015b) et ont fait l'objet de plusieurs présentations.

Sélection de groupes de variables par l'*adaptive ridge* Ce chapitre reprendra l'idée précédente mais dans le cadre de la sélection de groupes de variables, ce qui nécessite une définition plus précise des questions statistiques et un protocole adapté. Ces travaux ne sont pas encore publiés mais offrent une approche originale pour répondre aux objectifs biologiques, où une réponse "moins" précises, mais plus adaptée au problème/données, est trouvée pour répondre avec plus de confiance.

Pour mettre à l'épreuve nos algorithmes et nous comparer à l'exis-

tant, nous nous sommes basés principalement sur des simulations et sur un jeu de données réelles en grande dimension se portant des patients séropositifs qui avait déjà été étudié (Dalmasso et al. 2008). Ces différents jeux de données seront analysés sur l'ensemble des chapitres afin de voir une gradation sur les apports et la pertinence de nos propositions.

1. État de l'art

1.1 Sélection de variables pour la biologie et la médecine

La question de la sélection de variables en biologie et en médecine est un problème central que l'on soit dans des approches exploratoires ou prédictive. De manière exploratoire, on peut chercher les gènes ou les réseaux de régulations qui s'inscrivent dans des phénomènes précis comme la réponse à des traitements (Snouber et al. 2012) ou à des pathologies comme le virus du sida (Dalmasso et al. 2008). D'un point de vue plus médical, on pourra chercher les marqueurs d'un terrain à risque ou ceux permettant de choisir le traitement le plus efficace en fonction du profil du patient. Dans tous ces cas et surtout dans le dernier, choisir des cibles d'intérêts ne peut se faire sans un contrôle de l'incertitude.

Par exemple, l'étude d'association pangénomique (GWAS) de Dalmasso et al. (2008) cherche à retrouver les marqueurs génétiques associés à la charge virale du sida dans les cellules et le plasma. Cette analyse et les données associées sont typiques des problématiques de la grande dimension et de la manière dont les biologistes les analysent à l'heure actuelle. Les marqueurs génétiques utilisés sont les SNPs (*Single Nucleotides Polymorphisms*), qui représentent des variations sur le code génétique. On parle de SNP quand la variation pour un nucléotide est supérieure à 1% de la population. Ces marqueurs sont très pratiques car il sont placés de manière assez homogène le long du génome, avec un SNP, chez l'humain, toutes les centaines de bases en-

virons. La spécificité de ces variations génétiques dans la population et le fait qu'elles soient présentes en tout point du génome font qu'elles sont utilisées communément pour différencier des sous-populations ou des individus (elles représentent 90% des différences entre individus au niveau du génome). Pour cette analyse, le but est de déterminer les SNPs qui peuvent expliquer un différent degré de charge virale chez des personnes atteintes du sida parmi les 317,139 SNPs mesurés. Ceci permettrait de trouver des régions du génome impliquées dans la réponse à ce virus et également des empreintes génétiques permettant de pouvoir estimer comment un patient séropositif répondra au virus. La cohorte sur laquelle l'analyse a été faite ne comporte que 605 personnes pour 20,811 SNPs à analyser si nous ne nous intéressons qu'au SNPs du chromosome 6, qui semble le chromosome le plus intéressant pour cette étude selon la bibliographie. Le jeu de données associé sera présenté plus précisément dans le chapitre 3, quand nous y appliquerons notre procédure.

Nous faisons ainsi face au problème de la grande dimension où des variables comme des gènes ou des marqueurs génétiques sont en très grand nombre et les instances/exemples ou les individus des cohortes sont en nombre beaucoup plus faible. Pour les SNPs, se posent aussi des questions sur notre capacité à les distinguer en présence de fortes corrélations, que ce soit pour des raisons biologiques ou des raisons spatiales liées à leurs proximités sur le génome.

Nous sommes ainsi confrontés au problème de grande dimension. Il en découle deux problèmes, le premier est de pouvoir estimer les dépendances entres variables dans ce contexte, et le second est d'assortir cette estimation d'une mesure de confiance pour interpréter la significativité des dépendances estimées.

Nous considérerons dans ce manuscrit que l'on cherche à ajuster des modèles linéaires de la forme

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

où la matrice $\mathbf{X}$, de dimensions $n \times p$, correspond aux mesures de p

variables sur n échantillons, $\mathbf{y}$ est le vecteur réponse de dimension n, $\boldsymbol{\beta}$ est le vecteur, supposé parcimonieux, des coefficients inconnus sur les p variables et $\boldsymbol{\epsilon}$ est le terme d'erreur. Nous ne considérerons pas d'intercept β_0 car nous supposerons que les données sont centrées. En sélection, le but sera de retrouver l'ensemble des variables explicatives dont les coefficients de $\boldsymbol{\beta}$ sont non-nuls, qui constitueront le support $\mathcal{S}^\star = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$.

On résume souvent les problèmes de grande dimension aux problèmes $p \gg n$, mais il faut également tenir compte du bruit, ainsi que des corrélations entre variables. Outre le ratio p/n, il faut également considérer le ratio $|\mathcal{S}^\star|/n$, qui correspond au nombre de variables effectives par rapport au nombre d'exemples. Plus ces ratios sont petits et plus il sera facile de retrouver $\mathcal{S}^\star$. Dans la réalité, nous ne connaissons pas $\mathcal{S}^\star$ mais nous faisons l'hypothèse que le modèle est très parcimonieux et donc que $|\mathcal{S}^\star| \ll p$ et que $|\mathcal{S}^\star| < n$.

Estimer des paramètres en grande dimension est possible, comme nous le verrons par la suite, mais ce n'est qu'une étape dans un protocole de sélection de variables potentiellement explicatives. Tout biologiste ou analyste doit formuler, avant même de mettre en place son protocole, la question à laquelle il souhaite répondre. Ce ne sont pas les données qui posent les questions, ce sont les données qui doivent y répondre. Ainsi avant même de parler de quelle manière on peut estimer avec ou sans biais les paramètres inconnus $\boldsymbol{\beta}$, il parait plus logique de présenter les questions auxquelles on souhaite répondre et les tests statistiques qui y sont associés.

1.1.1 Test d'hypothèse

Dans des applications biologiques ou médicales, la sélection de variables devrait idéalement être sans erreur. À défaut, le risque d'erreur doit être contrôlé. Prendre la décision de sélectionner ou non une variable, c'est tester sa significativité. Pour construire ce test il faut définir les hypothèses sous-jacentes. Cela consiste à définir une hypothèse

TABLE 1.1 – Résultats possibles d'un test. Les lignes représentent la véracité des hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$, et les colonnes représentent la décision prise à l'issue test.

		DÉCISION	
		$\mathcal{H}_0$	$\mathcal{H}_1$
RÉALITÉ	$\mathcal{H}_0$	VN	FP
	$\mathcal{H}_1$	FN	VP

nulle, notée $\mathcal{H}_0$, qui corresponds à l'hypothèse prudente d'absence de significativité de la variable testée, contre une hypothèse alternative notée $\mathcal{H}_1$ correspondant à un état où la variable est déclarée comme significative. N'ayant aucun *a priori* sur l'hypothèse alternative $\mathcal{H}_1$, la décision se fait sur le rejet ou non de l'hypothèse nulle $\mathcal{H}_0$. Cette prise de décision entraîne quatre situations possibles représentées sur la table 1.1. Si $\mathcal{H}_0$ est vraie, le résultat du test est qualifié de Vrai Négatif (VN) ou de Faux Positif (FP) selon la décision prise, alors que si $\mathcal{H}_1$ est vraie, le résultat du test est soit qualifié de Vrai Positif (VP), soit de Faux Negatif (FN).

L'erreur de première espèce ou erreur de type I, notée α correspond à la probabilité de faux positifs (*false positive rate*, FPR) défini comme suit :

$$\text{FPR} = \mathbb{E}\left[\frac{FP}{FP+VN} I_{\{(FP+VN)>0\}}\right]. \tag{1.2}$$

Si en revanche $\mathcal{H}_0$ n'est pas rejetée à tort, on qualifie le résultat du test de faux négatif (FN), auquel est associé le risque de seconde espèce, ou erreur de type II, notée β (*false negative rate*, FNR). On préfère souvent utiliser le critère de puissance ou de sensibilité (SEN), défini comme suit :

$$\text{SEN} = \mathbb{E}\left[\frac{VP}{VP+FN} I_{\{(VP+FN)>0\}}\right]. \tag{1.3}$$

L'objectif est d'assurer un contrôle de l'erreur de type I fait en minimisant l'erreur de type II pour la sélection de variables en grande dimension. L'hypothèse $\mathcal{H}_0$ est alors celle d'absence d'effet d'une va-

riable ou d'un groupe de variables, et donc la nullité des coefficients de $\boldsymbol{\beta}$ associés.

Test de Student Dans le modèle linéaire (1.1), la variable j n'a aucune incidence si le coefficient β_j associé est nul. On peut donc définir les hypothèses comme suit :

$\mathcal{H}_0$: $\beta_j = 0$, la variable j n'est pas linéairement associée à $\mathbf{y}$

$\mathcal{H}_1$: $\beta_j \neq 0$, la variable j est linéairement associée à $\mathbf{y}$.

Le test de Student peut être utilisé pour tester l'hypothèse nulle si le terme d'erreur du modèle (1.1) est gaussien, en utilisant $\hat{\beta}_j$, l'estimateur des moindres carrés de β_j (nous reviendrons au chapitre 2 sur l'applicabilité de ce test en régression pénalisée). Dans ce contexte, la statistique de Student s'écrit

$$t_j = \frac{\hat{\beta}_j - \mathbb{E}(\hat{\beta}_j)}{\sqrt{s^2(\hat{\beta}_j)}} \ , \tag{1.4}$$

où $s^2(\hat{\beta}_j)$ est l'estimateur sans biais usuel de la variance de $\hat{\beta}_j$ et $\mathbb{E}(\hat{\beta}_j)$ son espérance.

Sous $\mathcal{H}_0$, la statistique t_j suit une loi de Student à $n - p$ degrés de liberté, loi qui se rapproche de la loi normale quand le nombre de degrés de liberté augmente comme l'illustre la figure 1.1. L'hypothèse nulle est rejetée si $t_j \in \left]-\infty, -s_{1-\alpha/2}\right] \cup \left[s_{1-\alpha/2}, \infty\right[$ où $s_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi de Student. Le test de Student est bilatéral : le risque est partagé entre les deux régions extrêmes pour que le risque global soit au niveau α.

Test de Fisher Le test de Fisher appliqué à la régression linéaire consiste à tester l'apport de variables pour la prédiction du vecteur réponse $\mathbf{y}$ au moyen de deux modèles emboîtés. Ces deux modèles sont respectivement le modèle complet Ω qui correspond à l'ensemble des variables et le modèle réduit $\boldsymbol{\omega}$ qui correspond à Ω duquel a été retiré la variable ou le groupe de variables que l'on souhaite tester (qui appartiennent donc au complémentaire $\bar{\boldsymbol{\omega}}$ de $\boldsymbol{\omega}$). Pour le test d'une seule variable, les hypothèses seront les mêmes que pour le test de

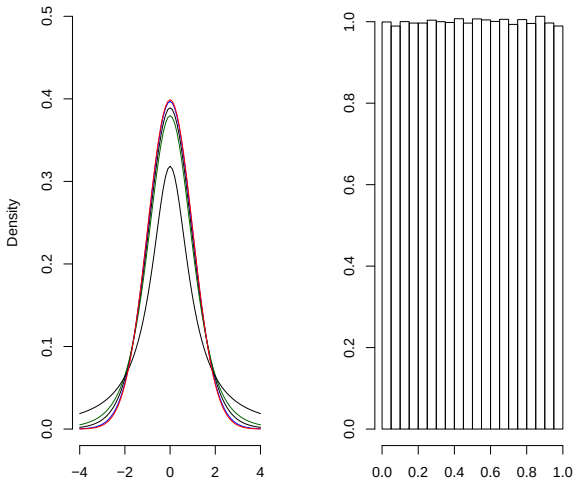


FIGURE 1.1 – À gauche : densités de la loi normale $\mathcal{N}(0,1)$ (en rouge), et de la loi de Student pour différents degrés de liberté (1, 5, 10 et 50). À droite : histogramme des p-valeurs sur 1000000 tirages issus d'une loi de Student.

Student ci-dessus. En généralisant à un ensemble arbitraire $\boldsymbol{\omega}$ on défini les hypothèses comme suit :

$\mathcal{H}_0 : \forall j \in \bar{\boldsymbol{\omega}} : \beta_j = 0$

$\mathcal{H}_1 : \exists j \in \bar{\boldsymbol{\omega}}$ tel que $\beta_j \neq 0$

La comparaison des deux modèles se fait sur l'erreur d'ajustement, en utilisant la somme des carrés des résidus (RSS) de chaque modèle. Ainsi l'hypothèse nulle se traduit par $\mathbb{E}\left[\mathrm{RSS}^{\omega} - \mathrm{RSS}^{\Omega}\right] = 0$. Contrairement au test de Student, le test de Fisher ne mesure donc pas directement le poids des variables testées, mais il en mesure l'importance au travers de leurs contributions à l'ajustement au vecteur réponse. La statistique de Fisher s'écrit :

$$F_{\bar{\boldsymbol{\omega}}} = \frac{\mathrm{RSS}^{\omega} - \mathrm{RSS}^{\Omega}}{\mathrm{RSS}^{\Omega}} \frac{\mathrm{df2}}{\mathrm{df1}} \quad . \tag{1.5}$$

Sous $\mathcal{H}_0$, elle suit la distribution de Fisher $\mathcal{F}_{\mathrm{df1,df2}}$, où df2 et df1 sont respectivement les degrés de liberté du modèle complet et la différence des degrés de liberté des deux modèles. Ainsi usuellement dans le cadre du test de Fisher en régression on définira $\mathrm{df1} = |\bar{\boldsymbol{\omega}}|$ et $\mathrm{df2} = n - p$ où $|\bar{\boldsymbol{\omega}}|$ correspond à la taille de l'ensemble des variables que l'on souhaite tester. La figure 1.2 représente la distribution de Fisher avec différents

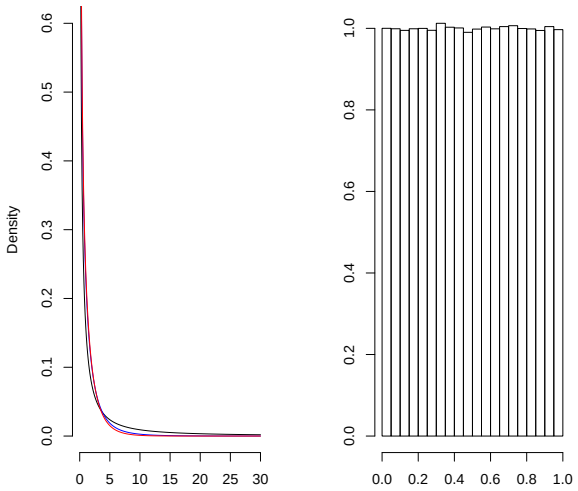


FIGURE 1.2 – À gauche : densité de lois de Fisher de différents couples de degrés de liberté. À droite : histogramme des p-valeurs sur 1000000 tirages issus d'une loi de Fisher.

paramètres. Le test est unilatéral, avec comme zone de rejet la queue de la distribution.

En régression l'application la plus commune du test de Fisher est de tester chaque variable séparément ainsi le modèle réduit $\boldsymbol{\omega}$ correspondra au modèle Ω sans la variable j que l'on souhaite tester. Dans ce cas, pour la régression OLS, le test de Fisher et le test de Student sont identiques. Il n'y a pas d'équivalence quand les estimateurs sont biaisés ou quand $|\bar{\boldsymbol{\omega}}| \neq 1$, ce qui sera le cas pour les groupes de variables étudiés au chapitre 4. Quand les estimateurs sont biaisés, la statistique de Fisher est moins impactée que la statistique de Student, ce qui justifie le choix de ce test pour mesurer l'importance des variables.

P-valeurs Un test d'hypothèse fait donc un choix binaire sur l'hypothèse $\mathcal{H}_0$. Classiquement, cette décision est assortie d'une p-valeur, notée P, qui est la probabilité d'observer une valeur au moins aussi extrême ou atypique de la statistique de test sous l'hypothèse nulle. Plus la p-valeur est faible, plus l'évènement est rare selon la loi supposée sous l'hypothèse nulle. Généralement, et dans tout les cas abordés

dans ce manuscrit, les p-valeurs des tests sous $\mathcal{H}_0$ suivront une distribution $\mathcal{U} \in]0,1[$, comme représentée sur les figures 1.1 et 1.2. Si l'on souhaite contrôler le risque de première espèce à un niveau α il suffit donc de rejeter $\mathcal{H}_0$ si

$$P \leq \alpha \ .$$

Dans le cadre de ce manuscrit nous nous sommes intéressés aux tests de Student et de Fisher qui nous semblaient les plus appropriés en régression linéaire.

Intervalles de confiance On préfère parfois l'estimation d'un intervalle de confiance à celle d'une p-valeur. Pour $\alpha \in [0,1]$ donné, un intervalle de confiance au niveau α a une probabilité de 1-α de contenir le vrai paramètre β_j. Dans le cas de la régression des moindres carrés, l'intervalle de confiance IC autour de la variable aléatoire $\hat{\beta}_j$ avec une confiance de niveau $1-\alpha$, tel que $\mathbb{P}(IC \ni \beta_j) = 1-\alpha$ est :

$$IC = \left[\hat{\beta}_j - q^{\mathcal{N}(0,1)}_{1-\alpha/2}\sqrt{\hat{\sigma}^2\left[\mathbf{X}^\top\mathbf{X}\right]^{-1}_{jj}}; \hat{\beta}_j + q^{\mathcal{N}(0,1)}_{1-\alpha/2}\sqrt{\hat{\sigma}^2\left[\mathbf{X}^\top\mathbf{X}\right]^{-1}_{jj}}\right] \quad (1.6)$$

où $q^{\mathcal{N}(0,1)}_{1-\alpha/2}$ est le $1-\alpha/2$ quantile de la loi normale centrée réduite et $\hat{\sigma}^2$ l'estimateur de la variance.

Les intervalles de confiance en plus de nous renseigner sur la significativité des variables apportent une information plus quantitative de l'effet de cette variable. Si l'hypothèse $\mathcal{H}_0$ correspond à la nullité de β_j et que 0 se trouve dans l'IC défini par $\hat{\beta}_j$ avec une confiance de $1-\alpha$, alors l'hypothèse nulle n'est pas rejetée au niveau α. Une p-valeur peut-être obtenue d'un intervalle de confiance mais l'intervalle de confiance se définissant par de l'estimateur $\hat{\beta}_j$ il faut que celui-ci ne soit pas biaisé ou que son biais soit connu. Dans le cadre de la sélection de variables, on préférera toutefois les p-valeurs qui permettent plus facilement de prendre en compte la multiplicité des tests, comme cela sera expliqué ci-dessous, même si des travaux similaires ont été menés pour les intervalles de confiance avec le *False Coverage Rate* (Benjamini and Yekutieli 2005).

1.1.2 Prise en compte de la multiplicité des tests

Les tests présentés ci-dessus permettent un contrôle de l'erreur de type I pour un test unique. Dans le cadre de données biologiques, le nombre de variables à tester est très important. Si on fait 10000 tests, en utilisant un seuil à 5%, alors on attendra en moyenne jusqu'à 500 faux positifs, ce qui correspondra bien à un FPR de 5%. De ce point de vue, interpréter les tests de manière unitaire sans prendre en compte leur multiplicité entraîne des problèmes d'interprétation. Le FPR est un critère fonctionnel dans le cas de tests simples mais il ne répond pas aux questions que l'on peut se poser dans le cadre des tests multiples. Effectivement quand m tests sont effectués il est préférable de contrôler, soit la probabilité d'avoir un FP parmi tous les tests, ou soit le taux de FP dans les tests rejetés. Ces deux critères sont respectivement le taux d'erreur sur l'ensemble des tests ($FWER$) et taux de fausses découvertes (FDR).

Contrôle du FWER Le taux d'erreur côté famille ($FWER$) est défini comme suit :

$$\text{FWER} = \mathbb{P}(FP \geq 1),$$

Ce critère mesure le risque de retrouver au moins un faux positif parmi tous les tests rejetés. Pour un test unique le FWER correspond donc à l'erreur de type I. Pour m tests indépendants, le $FWER$ sera égal, sur l'ensemble des tests, à $1-(1-\alpha)^m$ où α est le seuil de signification de chaque test. Le FWER sera donc supérieur à α dès que le nombre de tests sera supérieur à 1, même si pour chaque test l'erreur de type I est contrôlée à un risque α. La prise en compte des tests multiples est donc nécessaire si l'on souhaite contrôler le FWER à un niveau α sans inflation. La procédure la plus connue pour le contrôle du FWER est la procédure de Bonferroni. Elle consiste à modifier le seuil de rejet α, sur les p-valeurs, en tenant compte le nombre de tests. Ainsi un test

sera rejeté si

$$P_j \leq \frac{\alpha}{m}.$$

L'utilisation de cette procédure assure un contrôle du FWER au risque α. Malheureusement, cette procédure est très stricte et plus le nombre de tests est important plus il sera difficile de rejeter l'hypothèse nulle même quand elle devrait l'être.

Contrôle du FDR Benjamini and Hochberg (1995) ont proposé la procédure dite de Benjamini-Hochberg afin de contrôler le *False Discovery Rate* (FDR) au niveau de risque α. Le FDR s'exprime comme suit :

$$\text{FDR} = \mathbb{E}\left[\text{FDP}\right] \tag{1.7}$$

$$= \mathbb{E}\left[\frac{FP}{VP+FP}\mathbf{1}_{\{(VP+FP)>0\}}\right], \tag{1.8}$$

où FDP correspond à la proportion de fausse découverte pour une analyse. Le principal avantage de ce critère est qu'étant un taux, il est invariant à l'échelle, soit au nombre de tests. De plus il rejettera le plus grand nombre de tests tel que le critère soit contrôlé. La procédure de Benjamini-Hochberg consiste à chercher le rang h sur les m p-valeurs ordonnées de manières croissantes tel que $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ où

$$h = \max_{i\in\{1,\ldots,m\}}\{i : P_{(i)} \leq \frac{i\alpha}{m}\}. \tag{1.9}$$

Les tests qui sont rejetés par la procédure de Benjamini-Hochberg correspondent à toutes les variables de rang $i \leq h$, tel que représenté sur la figure 1.3 où la droite rouge représente la fonction de seuil $\frac{i\alpha}{m}$ et la valeur h représente le dernier test indexé sur i où la p-valeur est inférieure au seuil α. Le fait d'appliquer cette procédure assure un contrôle du FDR, mais le FDP peut avoir une grande variabilité qui doit être prise en compte pour l'interprétation, et qui sera discutée dans le chapitre 3.

En réalité, la procédure de Benjamini-Hochberg contrôle le FDR tel que

$$FDR = \pi_0\alpha \leq \alpha, \tag{1.10}$$

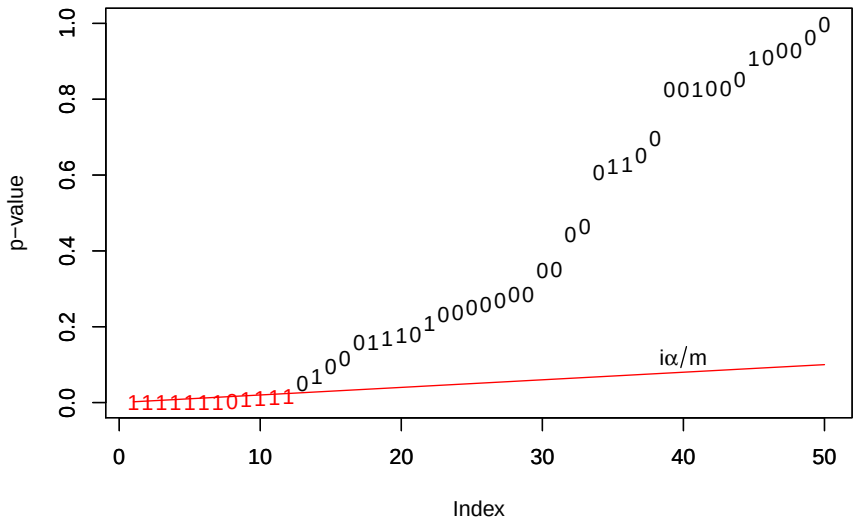


FIGURE 1.3 – Représentation de la procédure de Benjamini et Hochberg. La ligne rouge représente la fonction $i\alpha/m$ pour $\alpha = 0.1$ et $m = 50$. Les p-valeurs sont représentées par 1 si elles sont générées sous $\mathcal{H}_1$ et par 0 sous $\mathcal{H}_0$. Elle sont représentées en rouge si l'hypothèse nulle est rejetée car leur rang est inférieur ou égal à h calculé selon (1.9).

où π_0 est la proportion de tests pour lesquels l'hypothèse nulle est vraie. Cette proportion est inconnue mais est communément supposée proche 1. Toutefois il est possible que π_0 soit significativement plus petit que 1 ce qui va entraîner un contrôle plus sévère que celui attendu. Des méthodes comme QVALUE (Storey 2003) ou LBE (Dalmasso et al. 2005) visent à estimer π_0 de manière à assurer un contrôle effectif du FDR au niveau α choisi.

1.1.3 Procédures de test spécifiques à la sélection de variables

Critères sur des rangs Le contrôle du FDR par la procédure de Benjamini-Hochberg n'est pas toujours possible, par exemple quand nous n'avons pas accès à des p-valeurs mais simplement à des scores d'importance. Dans ce cas, les tests peuvent être ordonnés selon ce

score. Diverses procédures recherchent le rang maximum permettant de contrôler les faux positifs à l'instar du FDR. En régression, la manière la plus simple pour ranger les variables est de regarder les corrélations marginales entre les colonnes de $\mathbf{X}$ et la réponse $\mathbf{y}$. Des algorithmes existent pour générer les rangs basés sur le *Random Forest* (Breiman 2001),le *Support Vector Machine* (SVM)(Guyon et al. 2002, Rakotomamonjy 2003) ou plus générique comme l'algorithme ROGER (Jong et al. 2004). Des procédures basées sur les permutations peuvent être adaptées à ces méthodes en permettant le contrôle de critères proches du FDR comme le $pFDR$ (Listgarten and Heckerman 2007), le $eFDR$ (Ge et al. 2008) et le CER (Huynh-Thu et al. 2012). Ces algorithmes utilisent des permutations des données dans $\mathbf{X}$ selon le rang testé ou des permutations du vecteur réponse $\mathbf{y}$ dans le cas du $pFDR$. L'idée générale est toujours de mesurer l'effet de ces permutations sur le rang des scores estimés et de décider quel est le rang maximum où le signal permuté ne peut être confondu avec celui de référence en prenant un compte un niveau de risque α.

Notion de variables pertinentes Les critères présentés ci-dessus sont contrôlés au taux attendus quand les variables sont indépendantes mais cela n'est plus forcément vrai lorsque les variables sont fortement corrélées. Par exemple, la procédure de Benjamini-Hochberg aura un effet de sur-contrôle du FDR en cas de corrélations positives, le FDR sera donc strictement inférieur à α (Benjamini and Yekutieli 2001). Si plusieurs variables sont corrélées, elles auront un comportement se ressemblant et donc une spécificité amoindrie. Cela en résulte un effet sur les procédures de contrôle des tests et une instabilité dans le cadre des régression parcimonieuses comme le Lasso, que nous verrons plus tard.

La notion de variables pertinente peut être questionnée quand les variables sont structurées en groupes fortement corrélés. En effet doit-on représenter l'ensemble des variables du bloc comme explicatives, ou doit-on se contenter de définir comme seuls vrais positifs les va-

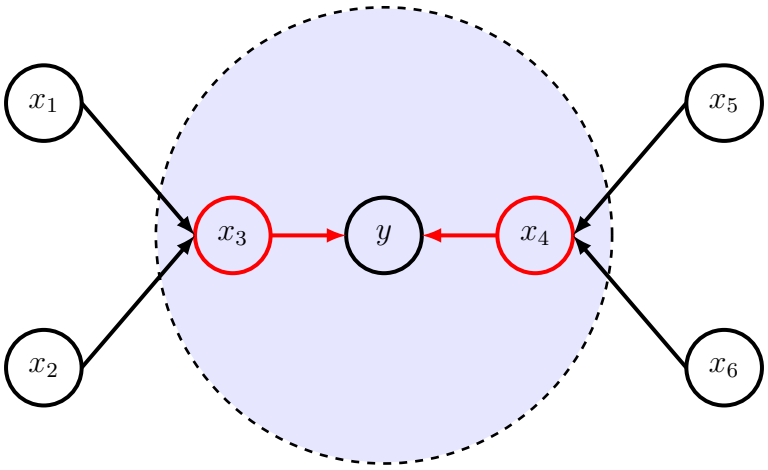


FIGURE 1.4 – **Représentation schématique de la couverture de Markov pour la variable de réponse y.** Les variables entourées en rouges x_3 et x_4 représentent les variables explicatives qui sont dans la couverture de Markov de y représentée par le disque grisé. Les autres variables sont corrélées aux variables explicatives mais ne sont pas dans la couverture et sont donc non-explicatives.

riables dont les coefficients β_j sont non nuls ? La figure 1.4 montre la couverture de Markov de la variable aléatoire y dans un problème de régression. Cette représentation induit le fait que seuls les parents de y sont suffisants pour prédire son comportement. Le choix de la couverture de Markov pour représenter notre problème définit que seules les variables directement explicatives doivent être définies comme étant l'ensemble des vrais positifs.

La figure 1.5 montre le changement d'état des variables dans la représentation de la couverture Markov quand une variable parente n'est plus observée. Suivant la couverture de Markov la variable x_1 pourra être prédite par ses variables parentes x_2 et x_4. Ainsi, l'ensemble de ces variables vont devenir pertinentes pour expliquer y. Dans notre cas, les variables parentes de x_1 correspondent aux variables qui lui sont corrélées. Cette représentation de l'état de pertinence des variables est transparente en pratique car nous connaissons rarement la réalité. En effet, nous observons un nombre limité de variables et ne pouvons savoir si une variable est explicative par nature ou par le fait qu'elle compense une variable non observée. Toutefois cette représentation montre que

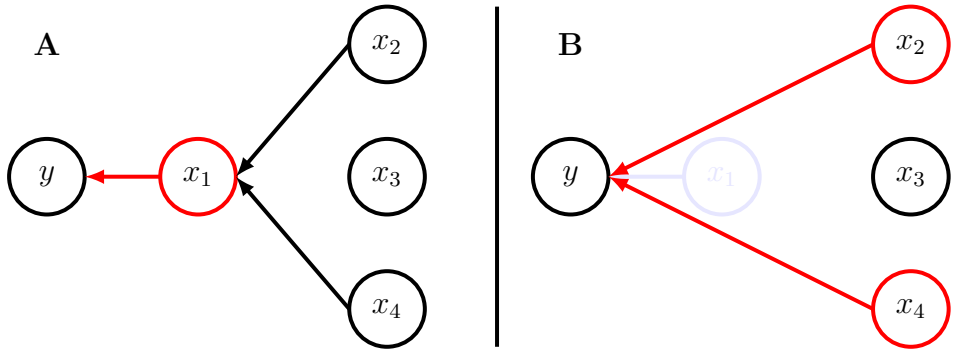


FIGURE 1.5 – Changement d'état des variables selon la représentation de la couverture de Markov de la variable $y = \beta x_1$. Les variables entourées en rouges représentent les variables explicatives selon les cas. Le cas A représente la couverture de y quand la variables explicative x_1 est observée et le cas B quand elle ne l'est pas où x_2 et x_4 deviennent pertinentes pour expliquer y en remplacement de x_1.

la notion de pertinence d'une variable dépend non seulement de sa capacité à expliquer notre phénomène mais également de la manière dont nous appréhendons la structure des variables.

La couverture de Markov est une représentation très stricte de la pertinence des variables, en effet cela suppose que les variables soient distinguables les unes des autres. Dans le cas où des variables sont trop fortement corrélées, il sera difficile de percevoir la spécificité de chacune et donc de les distinguer. Dans ce cas, la définition de la couverture de Markov ne sera pas forcément la plus pertinente pour représenter le caractère explicatif ou non d'une variable. Ainsi, le variables explicatives deviendraient l'ensemble des variables dans un groupe portant du signal et non plus les seules variables qui portent le signal dans le modèle d'origine.

Cela entraîne deux possibilités, soit les variables ne sont plus considérées en tant qu'éléments individuels mais en groupes pour les tests statistiques, soit toutes variables suffisamment corrélées à une variable explicative seront considérées comme tels. Dans ce sens, Yi et al. (2015) propose d'utiliser une mesure adaptée du FDR. Cette mesure, nommée $tFDR$ (Empirical thresholded false discovery rate) propose de définir un seuil de corrélation à partir duquel toutes les variables d'un groupe

sont soit considérées explicatives si au moins l'une d'entre elles l'est dans le modèle initial.

De manière moins évidente l'importance du bruit jouera également dans la capacité à distinguer la spécificité des variables corrélées, car si ce bruit est trop important alors ce problème deviendra excessivement difficile. Ce sujet sera abordé dans le cadre de la régression Lasso en 1.2.3.

1.2 Dans le cadre de la grande dimension

La section précédente a montré la nécessité de développer des approches statistiques pour la sélection de variables dans des applications biologiques ou médicales. Ces approches doivent, d'ailleurs, prendre en compte la notion de tests multiples pour contrôler le risque de faux positifs. Dans cette section, seront présentées les différentes méthodes de régression permettant d'estimer les coefficients $\boldsymbol{\beta}$ afin d'expliquer $\mathbf{y}$. Nous y discuterons également des tests qui existent sur ces estimateurs, sachant que ces méthodes seront confrontées à la problématique de la grande dimension. Cette présentation non exhaustive permettra d'avoir une vue d'ensemble de toutes les approches qui seront discutées tout au long de ce manuscrit.

1.2.1 Régression des moindres carrés

La régression ordinaire des moindres carrés (OLS) consiste à estimer de manière non biaisée les paramètres inconnus $\boldsymbol{\beta}$ par $\hat{\boldsymbol{\beta}}(0)$, minimisent $\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$. La solution de l'OLS s'écrit comme suit :

$$\hat{\boldsymbol{\beta}}(0) = (\mathbf{X}^\top \mathbf{X})^{-1} X^\top \mathbf{y}. \tag{1.11}$$

L'estimateur de l'OLS possède de nombreuses propriétés intéressantes. En effet les estimateurs $\hat{\boldsymbol{\beta}}(0)$ sont non biaisés et les statistiques qui peuvent être construites sur ces estimateurs ont des distributions connues dont les paramètres sont facilement estimables. Malheureusement, dans le cadre de la grande dimension, cette approche est in-

applicable. Effectivement si $n < p$ alors $\mathbf{X}^\top\mathbf{X}$ ne sera plus inversible, de même s'il existe au moins deux variables proportionnelles dans $\mathbf{X}$ comme dans le cas des variants rares, par exemple.

Approche univariée La régression marginale apparaît comme la plus simple manière de pouvoir estimer et tester la significativité d'une variable, sans tenir compte du problème de la grande dimension. C'est un cas particulier de l'OLS où l'on fait l'hypothèse que $\mathbf{X}^\top\mathbf{X} = I$ et donc cela revient à estimer les variables indépendamment des autres. Si $\sigma^2_{\mathbf{X}} = 1$, alors l'estimation des paramètres est définie comme suit :

$$\hat{\boldsymbol{\beta}}^{\text{univar}} = \frac{\mathbf{X}^\top\mathbf{y}}{n-1}. \tag{1.12}$$

Ainsi quel que soit le rapport n/p et le degré de colinéarité on pourra obtenir un estimateur de $cov(X_j, \mathbf{y})$. La significativité de cet estimateur sera mesurée par un test de Student. Évidemment cela revient à estimer un modèle sans tenir compte de l'interaction et surtout de la spécificité de chaque variable. Cette méthode est couramment utilisée en bio-informatique pour faire de la sélection de variables (Dalmasso et al. 2008) , pour sa simplicité, et sa facilité d'interprétation. Cette mesure peut également servir de critère de d'ordonnancement des variables pour des estimations de sous-ensembles optimaux (Wasserman and Roeder 2009).

1.2.2 Régression ridge

La régression ridge (Hoerl and Kennard 1970) est une méthode de régression pénalisée en norme $l2$ où les coefficients $\hat{\beta}_j(\lambda_2)$ sont ajustés tel que :

$$\hat{\beta}_j(\lambda_2) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, J(\boldsymbol{\beta}) + \lambda_2 \left\|\boldsymbol{\beta}\right\|_2^2, \tag{1.13}$$

où $\lambda_2 \in \mathbb{R}+$. La solution de cette équation, dans le cas où $J(\boldsymbol{\beta})$ correspond aux moindres carrés, s'écrit comme suit

$$\hat{\beta}_j(\lambda_2) = (\mathbf{X}^\top\mathbf{X} + \lambda_2\mathbf{I}_p)^{-1}X^\top\mathbf{y}. \tag{1.14}$$

L'utilisation de la pénalisation en norme $l2$ permet de contracter les estimateurs $\hat{\boldsymbol{\beta}}$ des coefficients $\boldsymbol{\beta}$. Plus le facteur de pénalité λ_2 tend vers $+\infty$, plus les coefficients $\hat{\beta}_j(\lambda_2)$ tendent vers 0, et si cette pénalité est nulle on retrouve les estimateurs de l'OLS. Le choix de λ_2 peut s'apprendre directement sur les données, mais cela n'est possible qu'en petite dimension car les définitions existantes nécessites les estimateurs de moindres carrés. Dans le cas de la grande dimension, le choix de λ_2 se fait communément par validation croisée. La régression ridge ne permet donc pas une sélection directe de variables telle que peut le faire la régression en norme $l1$ nommé Lasso (Tibshirani 1996). Toutefois la régression ridge permet de mieux gérer le problème de multicolinéarité par rapport à l'OLS. Elle permet surtout de pouvoir être résolue quand $n < p$ car la matrice $\mathbf{X}^\top\mathbf{X} + \lambda_2\mathbf{I}_p$ est toujours inversible si $\lambda_2 > 0$.

Les travaux de Halawa and El Bassiouni (1999) se sont portés sur l'utilisation du test de Student pour tester la significativité des coefficients, mais ces résultats théoriques sont difficilement applicables en réalité. En effet, le fait que les coefficients de la ridge soient biaisés par la pénalisation font, entre autres, qu'il n'est pas possible de mettre en place des tests statistiques, comme le test de Student ou de Fisher en régression, sans corriger ce problème de biais. L'étude approfondie de la régression ridge et de la possibilité d'effectuer des tests d'hypothèse à partir ses estimateurs sera le sujet du chapitre 2.

1.2.3 Régression Lasso

La procédure Lasso (Least Absolute Shrinkage and Selection Operator (Tibshirani 1996) est une méthode de régression faisant intervenir une contrainte sur la norme $l1$ des coefficients. L'estimateur $\hat{\boldsymbol{\beta}}$ est défini par :

$$\hat{\boldsymbol{\beta}}(\lambda_1) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, J(\boldsymbol{\beta}) + \lambda_1 \left\|\boldsymbol{\beta}\right\|_1 \quad , \tag{1.15}$$

où $\lambda_1 \in \mathbb{R}+$.

Contrairement à la régression ridge, il n'y a pas de solution analytique pour le Lasso, mais ce problème étant convexe il peut être résolu

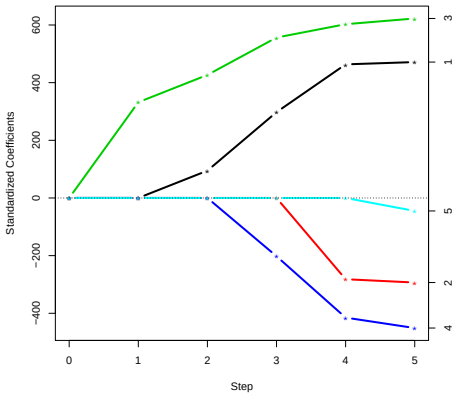


FIGURE 1.6 – Chemin de régularisation de l'algorithme LARS.

"facilement" de manière exacte. Le Lasso est très apprécié en sélection car il est parcimonieux, c'est à dire qu'il permet d'éteindre un certain nombre de variables en leur affectant un coefficient nul dans $\hat{\boldsymbol{\beta}}(\lambda_1)$. Plus le paramètre λ_1 est grand, plus le modèle sera parcimonieux, ainsi pour λ_1 grand tous les coefficients de $\hat{\boldsymbol{\beta}}(\lambda_1)$ seront nuls. Par la suite nous noterons l'ensemble :

$$\hat{\mathcal{S}}(\lambda_1) = \{j : \hat{\beta}_j(\lambda_1) \neq 0\}$$

correspondant aux variables non-nulles pour une pénalité λ_1 donnée et $|\hat{\mathcal{S}}(\lambda_1)|$ la taille de cet ensemble.

Différents algorithmes existent pour résoudre le problème (1.15), mais le plus connu est la procédure itérative LARS (Efron et al. 2004). Elle consiste à faire décroitre la pénalité λ_1 dans l'intervalle $]0, +\infty[$ en recherchant les pénalités pour lesquelles une variable $j \in 1, ..., p$ pourra passer d'un coefficient nul à un coefficient non nul et inversement. Le chemin de régularisation de l'algorithme LARS, représenté dans la figure 1.6, montre l'entrée des différentes variables en fonction de la pénalité λ_1. Dans ce chemin de régularisation nous pouvons voir qu'il y a deux manières d'appréhender les résultats du Lasso. La manière la plus simple est d'observer les coefficients estimés pour une valeur λ_1. Ainsi on pourra retenir le signe et le poids des différents coefficients afin de les ordonner et de les interpréter. L'autre manière consiste à ordonner les variables selon l'ordre d'entrée dans le chemin

de régularisation où les variables les plus pertinentes sont censées être sélectionnées en premières (Lockhart et al. 2014). Comme dans le cadre de la régression ridge, le choix de λ_1 se fait généralement par validation croisée.

Propriétés du Lasso Le Lasso offre un estimateur parcimonieux permettant une sélection directe d'un sous-ensemble de variables pour lesquelles $\hat{\beta}_j(\lambda_1) \neq 0$. En général, ce sous-ensemble ne pourra pas avoir une taille plus grande que n quelque soit la pénalité λ_1 non-nulle choisie. Ceci ne pose pas de problème si $n > |\mathcal{S}^\star|$, mais dans le cas de petits échantillons, ce qui est fréquent en analyse de biopuces, on risque d'avoir $n < |\mathcal{S}^\star|$ et donc que $\mathcal{S}^\star$ ne puisse pas être retrouvé dans le support estimé $\hat{\mathcal{S}}(\lambda_1)$. De plus pour ce type de données, p est typiquement très grand devant n ce qui fait que ce genre de problème est très difficile. La contrainte sur la taille de $\hat{\mathcal{S}}(\lambda_1)$ aura toutefois son utilité dans le cadre des travaux de Wasserman and Roeder (2009), qui seront abordés dans le chapitre 3 et cette contrainte sera nommée "screening property" (Meinshausen et al. 2009).

Pour autant, ce sous-ensemble $\hat{\mathcal{S}}(\lambda_1)$ est, sous des conditions sur lesquelles nous reviendrons plus tard, asymptotiquement consistant (Zhao and Yu 2006) :

$$\lim_{n\to\infty} \mathbb{P}(\mathcal{S}^\star \subseteq \hat{\mathcal{S}}(\lambda_1)) = 1, \tag{1.16}$$

ce qui signifie que les variables explicatives se retrouvent asymptotiquement dans l'ensemble $\hat{\mathcal{S}}(\lambda_1)$ sélectionné par le Lasso. Cette consistance assure également une cohérence sur le signe des coefficients. Il est intéressant de noter que $\mathcal{S}^\star$ a de grandes chances d'être inclus dans $\hat{\mathcal{S}}(\lambda_1)$ mais que nous ne pouvons y estimer le taux de faux positifs. Le support $\hat{\mathcal{S}}(\lambda_1)$ est donc difficilement interprétable sur l'effet réel de chaque variable. Cependant si la sélection de modèles se fait simplement dans un but de prédiction alors ce problème d'interprétation ne se pose pas ou moins.

Instabilité du Lasso La propriété de consistance exprimée ci-dessus, n'est établie que sous la condition d'irreprésentabilité (Zhao and Yu 2006). Ainsi, une variable j sera confondue avec une variable i, tant que $cov(x_j, x_i) > var(x_j)$, car le Lasso ne pourra distinguer l'effet spécifique de j si les deux variables ont un comportement trop proche. De ce fait l'estimateur Lasso n'est pas très sensible à la spécificité des variables, et si des variables sont fortement corrélées, le Lasso les sélectionnera de manière "aléatoire" si au moins une peut expliquer $\mathbf{y}$. Quand la condition d'irreprésentabilité n'est pas vérifiée, le Lasso ne permet plus d'assurer de retrouver $\mathcal{S}^\star$ dans $\hat{\mathcal{S}}(\lambda_1)$ si des variables explicatives sont confondues avec des non-explicatives. C'est en ce sens que le Lasso peut-être considéré comme une procédure instable. Pour gérer ce problème il existe plusieurs solutions : la plus simple serait de ne plus définir la sélection en termes de variables mais en termes de groupes de variables. Cela entraîne d'autres problèmes comme la définition des groupes, leur éventuel recouvrement, etc, sur lesquels nous reviendrons ultérieurement dans le chapitre 4.

Une autre approche nommée "stability selection" basée sur le rééchantillonnage, a été proposée par Meinshausen and Bühlmann (2010). Elle consiste à effectuer un nombre "suffisant" de Lasso sur des sous-échantillons aléatoires de taille $\frac{n}{2}$. Le sous-ensemble final $\hat{\mathcal{S}}(\lambda_1)$ est défini par les variables qui ont un taux de sélection supérieur à un seuil π sur l'ensemble des sous-échantillons. Cette approche partage le principe du *boLasso* (Bach 2008) qui, en théorie, cherche l'intersection entre les supports sélectionnés sur les sous-échantillons. Ces deux approches permettent, en théorie de sélectionner un support $\hat{\mathcal{S}}(\lambda_1)$ ayant un meilleur recouvrement de $\mathcal{S}^\star$ que le simple Lasso. Il n'y a également, à ce jour, pas de possibilité de contrôler le FDR même si dans le cas de la stability selection (Meinshausen and Bühlmann 2010) on peut contrôler le FWER. Toutefois, pour que cette dernière contrôle le FWER, cela nécessite l'estimation d'un paramètre en plus du seuil .

Une approche de rééchantillonage sur les échantillons ne suffit pas forcément à répondre au problème de variables hautement corrélées qui

peuvent toujours se masquer. Ainsi Beinrucker et al. (2014) proposent de ré-échantillonner en parallèle les échantillons et les variables afin de minimiser le masquage des variables entres elles.

Quantification de l'incertitude sur les estimateurs du Lasso

Les deux types de quantification de l'incertitude sur des coefficients de régression sont les intervalles de confiance et les p-valeurs (1.1.1). Pour du Lasso, il est très difficile de définir des statistiques sur la précision ou la nullité des estimateurs, du fait de l'utilisation d'une pénalité qui induit, entre autre, un estimateur biaisé de $\boldsymbol{\beta}$ (comme dans la régression ridge, cf 1.2.2).

Intervalles de confiance (IC) Chatterjee and Lahiri (2010) proposent d'estimer un IC sur les coefficients du Lasso par le biais du bootstrap des résidus (rééchantillonage avec remise). La distribution des coefficients issus de ce bootstrap permet de définir un IC par variable. Le Lasso étant très instable, il y est proposé de mettre un seuil a où tout $|\hat{\boldsymbol{\beta}}(\lambda_1)| \leq a$ sera considéré comme nul. Ce seuillage permettant de diminuer la taille de $\hat{\mathcal{S}}(\lambda_1)$ en rejetant les variables ayant un coefficient de trop faible poids. Il peut faire penser à la règle 1se (*One Standard Error*) (Breiman et al. 1984), qui s'applique dans un processus du choix de la pénalité optimale en validation croisée. Pour chaque valeurs de λ_1 on peut estimer un écart-type mesuré sur les erreurs d'estimation obtenues sur chaque sous-ensembles de validation. La règle 1se consiste à choisir la pénalité la plus grande parmi les valeurs de λ_1 pour lesquelles l'erreur d'estimation est à moins d'un écart type de l'erreur d'estimation de la valeur λ_1 initialement choisie. De ce fait le seuillage proposé par Chaterjee est moins interprétable que cette dernière. L'ajout de ce paramètre a nécessite donc d'optimiser le couple $\langle \lambda_1, a \rangle$ dans le cadre de la validation croisée. L'intervalle de confiance ainsi produit, présente une couverture proche de la couverture que l'on aurait pu obtenir par l'OLS (Sartori 2010).

Tests statistiques pour le Lasso De nombreuses recherches ont été faites afin de tester la pertinence des variables directement dans la procédure Lasso. Ainsi le test de covariance (Lockhart et al. 2014) a été construit sur la procédure LARS (Efron et al. 2004). La statistique de test pour la variable $j \in \{1, ..., p\}$ s'exprime comme suit :

$$t_j = \frac{1}{\sigma^2} \left(\langle \mathbf{y}, \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_{(j+1)}) \rangle - \langle \mathbf{y}, \mathbf{X}_{\hat{\mathcal{S}}}\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}(\lambda_{(j+1)}) \rangle \right) \quad , \tag{1.17}$$

où $\mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_{(j+1)})$ est l'estimation de $\mathbf{y}$ pour la pénalité $\lambda_{(j+1)}$, définie comme celle pour laquelle $\hat{\beta}_j$ est pour la première fois non nul sur le chemin de régularisation du LARS. Le second terme $\mathbf{X}_{\hat{\mathcal{S}}}\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}(\lambda_{(j+1)})$ reprend le même principe mais il est estimé sur l'ensemble de variables $\hat{\mathcal{S}}$ qui contient toutes les variables avec un coefficient non nul excepté la variable j que l'on souhaite tester.

En résumé, ce test pour une pénalité donnée (correspondant à l'entrée de la variable dans le chemin de régularisation) permet de voir si l'ajout de cette variable apporte une information significative pour l'estimation de $\mathbf{y}$. Malheureusement, cette approche nécessite de connaître la variance du terme d'erreur σ^2, qui peut être estimée facilement quand $n > p$ mais qui est difficilement estimable en grande dimension. On retrouve dans cette procédure de test le principe de comparaison de modèles emboîtés que nous utilisons dans le cadre de la régression ridge (cf équation 2.17). La principale différence de ces deux tests est que dans notre cas nous testons l'importance d'une variable par rapport à l'ensemble des autres alors que dans Lockhart et al. (2014) c'est son importance par rapport aux variables ayant un rang supérieur. Le principe de test séquentiel, où chaque test dépend des résultats du précédent, offre un intérêt tout particulièrement dans le cas de variables hautement corrélées. Ceci pourrait éviter en partie le problème de masquage de variables, qui peut arriver dans le cadre du test de la variables contre toutes les autres. En revanche, des tests séquentiels, donc dépendants, empêchent d'appliquer un contrôle du FDR.

Une approche naïve pour tester la significativité des estimateurs

du Lasso serait de transformer les données par permutation afin de simuler ces données sous l'hypothèse nulle testée. Toutefois dans le cadre du Lasso ce serait très couteux en temps de calcul et difficilement justifiable théoriquement (Chatterjee and Lahiri 2013).

Une alternative efficace, le knockoff test, a été proposée par Barber and Candes (2014) sur ce principe de génération de données sous l'hypothèse nulle sans utiliser de permutations. La première étape de cette procédure est de générer une matrice dite "knockoff" $\tilde{\mathbf{X}}$, qui correspond à une transformation de $\mathbf{X}$ où la structure de corrélation est conservée, ainsi $\mathbf{X}^\top\mathbf{X} = \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$. La matrice "knockoff" possède de nombreuses propriétés où $\mathbf{X}_j^\top\tilde{\mathbf{X}}_k = \mathbf{X}_j^\top\mathbf{X}_k$ pour chaque $j \neq k$, $\mathbf{X}_j^\top\tilde{\mathbf{X}}_j \neq 1$ et $\mathbf{X}_j^\top\mathbf{X}_j = 1$. Une variable $\tilde{j}$ dans $\tilde{\mathbf{X}}$ se comporte donc de la même manière par rapport aux autres variables qu'elles soient dans $\mathbf{X}$ ou $\tilde{\mathbf{X}}$, excepté par rapport à son homologue j. Ainsi les variables de la matrice "knockoff" sont suffisamment différentes de leurs homologues pour que l'hypothèse nulle soit vérifiée tout en n'altérant pas la structure des données. Ensuite, Barber and Candes (2014) propose d'estimer le Lasso sur la matrice augmentée $\left[\mathbf{X}\tilde{\mathbf{X}}\right]$ et de mesurer l'entrée préférentielle de j dans le chemin de régularisation par rapport à son homologue $\tilde{j}$ sous l'hypothèse nulle. Cette approche sur la stabilité des rangs analyse des critères proche du CER, présenté précédemment, mais sans permutation. La statistique de test obtenue pourra être contrôlée à un FDR de niveau α, où le seuil correspondant est appris sur les données. Cette approche offre l'avantage de préserver la structure de covariance quand une variable est testée, alors quelle est affectée dans le test que nous proposons en 2.17. Il n'est toutefois pas possible de construire ce test dans le cadre de la grande dimension car il se base sur l'inversion de la matrice $\mathbf{X}^\top\mathbf{X}$ pour générer $\tilde{\mathbf{X}}$, et cette matrice est singulière quand $n < p$. Il faut toutefois noter deux cas de figure, avec un cas favorable où $n > 2p$ et un cas moins favorable où $2p > n > p$. En effet il faut que la matrice augmentée ait une dimension $n > 2p$ pour pouvoir effectuer ce test, et si ce n'est pas le cas une transformation préalable est nécessaire, ce qui entraîne une augmentation du bruit.

Cette approche peut-être adaptée à la régression des moindres carrés et la régression ridge, en se basant sur les coefficients estimés et non sur le rang dans le chemin de régularisation pour le Lasso. Ceci sera discuté dans le chapitre 3.

Nous verrons en 3.1 des approches en deux étapes qui permettent de pouvoir estimer des p-valeurs sur la significativité de la sélection des variables dans $\hat{\mathcal{S}}(\lambda_1)$.

1.2.4 *Elastic-Net*

L'elastic-net (Zou and Hastie 2005) est une méthode de régression très proche du Lasso mais qui combine deux contraintes, une sur la norme $l1$ des coefficients et une sur leur norme $l2$. Le problème à résoudre est le suivant

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} J(\boldsymbol{\beta}) + \left(\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2\right) \quad ,$$

où λ_1 et λ_2 correspondent respectivement aux pénalités sur les normes $l1$ et $l2$ des coefficients. Ces deux paramètres permettent de donner plus ou moins de poids aux pénalités Lasso et ridge. En effet si $\lambda_1 = 0$ alors nous aurons l'estimateur ridge et si $\lambda_2 = 0$ nous retrouvons l'estimateur Lasso. L'elastic-net se révèle être plus souple et surtout plus stable que le Lasso tout en conservant son caractère parcimonieux.

La contrainte sur la norme $l2$ des coefficients permet d'éviter un trop grand masquage de variables explicatives quand elles sont corrélées, en favorisant un comportement similaire des coefficients pour ces variables corrélées. En théorie cela devrait permettre un meilleur recouvrement de $\mathcal{S}^\star$ par $\hat{\mathcal{S}}(\lambda_1)$ en présence de groupes de variables. Par contre, si les variables dans $\mathbf{X}$ sont indépendantes alors le Lasso et l'elastic-net devrait avoir un comportement équivalent. On peut noter que l'elastic-net, si il est parcimonieux, ne permet pas d'assurer que $|\hat{\mathcal{S}}| < n$ comme c'était le cas dans le cadre du Lasso. Ceci rendra incompatible l'elastic-net avec le protocole de Wasserman and Roeder (2009) qui sera présenté dans le chapitre 3.

1.2.5 Régression pénalisée et groupes de variables

Comme évoqué précédemment, la notion de variable pertinente peut-être redéfinie quand les variables sont corrélées. Une solution possible consiste à ne plus considérer les variables individuellement mais comme membres de groupes de variables. La définition des groupes, qui ne sera pas abordée dans ce manuscrit, peut découler de connaissances *a priori*, de l'utilisation d'algorithmes de classification de variables, ou de mesures empiriques des corrélations. L'appartenance d'une variable à un groupe pourrait ne pas être exclusive, mais nous ne considérerons ici que le cas où les groupes forment une partition de l'ensemble des variables. Que les groupes soient connus *a priori* ou appris sur les données, il existe différentes manières d'utiliser cette information pour questionner la pertinence du groupe, ce qui change la notion de faux et de vrais positif.

Cluster Representative Lasso La régression Lasso a été présentée comme très instable en présence de variables fortement corrélées. On peut dès lors considérer que la sélection d'une variable dans un groupe de variables corrélées est relativement arbitraire, et qu'elle correspond plutôt à sélectionner le groupe. Le *Cluster Representative Lasso* proposé par Bühlmann et al. (2013) consiste à construire une partition des variables en se basant sur leur corrélations. Les variables sont scindées en G groupes par un algorithme de classification ascendante hiérarchique. Les ensembles d'indices correspondants sont notés $\{\mathcal{G}_g\}_{g=1}^{G}$. De nouvelles variables « prototypes » sont définies en créant une matrice de pseudo-données $\tilde{\mathbf{X}}$, de taille $n \times G$, où les variables du groupe $\mathcal{G}_g$ ne seront plus représentées que par une seule variable :

$$\tilde{\mathbf{x}}_g = \frac{1}{|\mathcal{G}_g|} \sum_{j \in \mathcal{G}_g} \mathbf{x}_j \quad , \tag{1.18}$$

qui est la moyenne des variables $j \in \mathcal{G}_g$. Une fois la matrice $\tilde{\mathbf{X}}$ définie, il suffit d'y appliquer la régression Lasso sur l'espace des pseudo-données :

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^G} \tilde{J}(\boldsymbol{\beta}) + \lambda_1 \left\| \boldsymbol{\beta} \right\|_1 \quad .$$

Cette approche facilite l'interprétation des résultats, simplifie le problème et stabilise l'estimation. Toutefois, il en résulte une perte d'information qui peut-être dommageable, en particulier si il y a des corrélations négatives dans un groupe : deux variables peuvent alors se neutraliser.

Group-Lasso & sparse-group-Lasso Le *group-Lasso* prend en compte les groupes de variables directement à partir des variables d'origine (Yuan and Lin 2006). Il est défini comme suit :

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} J(\boldsymbol{\beta}) + \lambda_2 \sum_{g=1}^{G} \sqrt{w_g} \left\| \boldsymbol{\beta}_{\mathcal{G}_g} \right\|_2 \quad , \tag{1.19}$$

où $w_g = |\mathcal{G}_g|$ est le cardinal du groupe g, et $\boldsymbol{\beta}_{\mathcal{G}_g}$ représente le vecteur de dimension w_g formé des coefficients $\{\beta_j\}_{j\in\mathcal{G}_g}$.

Dans le *group-Lasso*, la norme euclidienne des coefficients de chaque groupe généralise la valeur absolue des coefficients utilisée dans le Lasso (1.15). L'utilisation de cette pénalité encourage toutes les coefficients d'un même groupe à être simultanément nuls. Contrairement au *Cluster Representative Lasso*, il n'y a pas besoin de transformer les données, mais le problème d'estimation reste en dimension p, et est par là plus délicat à résoudre.

Le *sparse-group-Lasso* (SGL, Friedman et al. 2010) est intermédiaire entre le *group-Lasso* et le Lasso :

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} J(\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{g=1}^{G} \sqrt{w_g} \left\| \boldsymbol{\beta}_{\mathcal{G}_g} \right\|_2 \quad , \tag{1.20}$$

où l'équilibre entre Lasso et *group-Lasso* est réglé par le ratio λ_1/λ_2. La pénalité en norme $l1$ ajoutée à celle du *group-Lasso* encourage une certaine parcimonie à l'intérieur des groupes sélectionnés.

Si ces deux dernières approches permettent de sélectionner conjointement des variables d'un groupe, elles n'assurent pas la cohérence du signe des coefficients au sein des groupes sélectionnés. Le Coop-Lasso (Chiquet et al. 2012) permet d'avoir un effet similaire au SGL tout en encourageant la cohérence du signe des coefficients des variables au

sein d'un groupe. Ceci peut permettre une meilleure interprétation des résultats.

2. Statistical Testing in Ridge Regression

We consider testing the relevance of the predictor variables in the high-dimensional linear regression setting, based on ridge regression estimates. Ridge regression is an ℓ_2 penalized estimator originally proposed for dealing with correlated design matrices. Its Tikhonov regularization allows for applications in high-dimensional settings, contrary to the ordinary least squares estimator which is then ill-defined. Most studies pertaining to variable selection from the ridge estimator consider settings with many more observations than covariates, leading to vanishing regularization parameters. However, the usual strategies for choosing the penalty parameter, either based on cross-validation or information criteria, often lead to sizable penalties in high-dimensional settings, leading to behaviors that are far from the small penalties that are suitable in the asymptotic regime. This chapter reviews the existing statistical tests pertaining to variable selection based on the ridge regression estimates, and shows that they perform poorly in high-dimension. This chapter introduces a permutation test which controls type I error in high-dimension and compares this new procedure to classical resampling testing strategies.

2.1 Introduction

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\epsilon} \ ,$$

where $\mathbf{y} = (y_1, \cdots, y_n)^t$ is the vector of responses, $\mathbf{X}$ is the $n \times p$ design matrix with $p \gg n$, $\boldsymbol{\beta}^\star$ is p-dimensional vector of unknown parameters, assumed to be sparse, and $\boldsymbol{\epsilon}$ is a n-dimensional vector of independent random variables of mean zero and variance σ^2.

This type of model dates back to Gauss and Legendre and still enjoys many ongoing practical and theoretical development two hundred years from its invention. At least two different families of problems can be tackled by means of this simple model: prediction and discovery/interpretation of the relationships between the response and the covariates. When predicting responses, $\boldsymbol{\beta}^\star$ is only the focus of attention as a tool to compute y_i. Then, the performances of an estimator $\hat{\boldsymbol{\beta}}$ are typically measured in terms of distance between the optimal linear predictor of y_i given $\mathbf{x_i}^t\boldsymbol{\beta}^\star$, and its estimation $\mathbf{x_i}^t\hat{\boldsymbol{\beta}}$. When focusing on the relationships between the response variable and the covariates, the problem may take two different forms: detection and selection. Detection answers the question of whether there is any significant signal in the covariates (is there at least one non-zero entry in the "true" parameter vector $\boldsymbol{\beta}^\star$), whereas selection considers the more specific question of finding which entries of $\boldsymbol{\beta}^\star$ are non-zero. A classical way to address the selection problem is statistical hypothesis testing where each coefficient of the arameter vector $\boldsymbol{\beta}^\star$ is tested for being null or not.

When the columns of the design matrix are correlated, estimating $\boldsymbol{\beta}^\star$ can be difficult. A common practice consists in dropping some of the covariates to improve the condition number of the design matrix, ideally forming a new design matrix whose covariance matrix is close to identity. A possible alternative consists in defining new variables which enjoy an (approximate) independence property. This may be achieved

by linear combinations of the original variables. Many combinations exist, such as simple averaging, principal component analysis, or partial least squares.

Another path for circumventing the correlated design problem is ridge regression (Hoerl and Kennard 1970). Ridge regression is a penalized version of the ordinary least squares (OLS) regression

$$\hat{\boldsymbol{\beta}}(\lambda_2) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \sum_{j=1}^{p} {\beta_j}^2 \ , \tag{2.1}$$

where λ_2 is a positive real penalty parameter. The solution to this problem is well-known

$$\hat{\boldsymbol{\beta}}(\lambda_2) = (\mathbf{X}^\top\mathbf{X} + \lambda_2\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} \ . \tag{2.2}$$

Ridge regression can always improve the mean square error of estimation and prediction. It also enjoys other advantages such as being suited for estimating regression parameters in the high-dimensional setting. The properties of ridge regression estimation are often related to the bias-variance trade-off. Indeed, contrary to OLS, ridge regression produces biased estimates, which owe their eventual advantage to a reduction in variance. Due to their bias, testing the nullity of the ridge regression coefficients is not as straitforward as for the classical OLS estimate : the exact distributions of the usual test statistics are not known.

Halawa and El Bassiouni (1999) proposed a modified Student's test to assess the significance of each covariate. Although this proposition is asymptotically justified, we argue that, in the $p > n$ context, the significance level of this test is not controlled, and that application papers using this approach (Cule et al. 2011) may produce grossly inaccurate results.

We propose here to investigate ridge regression estimation in the high-dimensional setting, focussing on significance testing of each variable. In a first section, the proposition of Halawa and El Bassiouni (1999) is examined and its limitations are exhibited. We then present

different resampling strategies for simulating samples from the distribution of the ridge estimates. Bootstrap and permutation strategies are indeed effective approaches for testing parameters when the null distribution is unknown. Their merits will be shown in the numerical experiments.

2.2 Ridge Regression in High-Dimension

Ridge regression offers several advantages over OLS. The estimate always exists for any $\lambda_2 > 0$, whereas the OLS problem is ill-defined when the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ is not of rank n, which is obviously the case when $p > n$.

Hoerl and Kennard (1970) also shown the so-called existence property about the Mean Square Error (MSE), which is defined as follow:

$$MSE(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\star})^2\right) \quad .$$

This property states that there always exists a penalty parameter $\lambda_2 > 0$ such that

$$MSE(\hat{\boldsymbol{\beta}}(\lambda_2)) < MSE(\hat{\boldsymbol{\beta}}(0)) \quad ,$$

where the ridge estimate $\hat{\boldsymbol{\beta}}(0)$, with $\lambda_2 = 0$, corresponds to the OLS estimate. The proof relies on the bias-variance trade-off. The variance of the ridge estimate decreases with λ_2 whereas its bias increases. The proof shows that there always exists a strictly positive value of λ_2 for which the gain in MSE due to variance reduction is greater than the loss due to squared bias (see Figure 2.1).

That being said, looking for a penalty parameter λ_2 minimizing the Mean Squared Error is an important concern when actually applying ridge regression. Crivelli et al. (1995) compare some strategies for setting λ_2, with

$$\lambda_2{}^{HOR} = \frac{p\,\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}(0)^{\top}\hat{\boldsymbol{\beta}}(0)}$$

and

$$\lambda_2{}^{LOR} = \frac{p\,\hat{\sigma}^2}{\hat{\boldsymbol{\beta}}(0)^{\top}\mathbf{X}^{\top}\mathbf{X}\hat{\boldsymbol{\beta}}(0)},$$

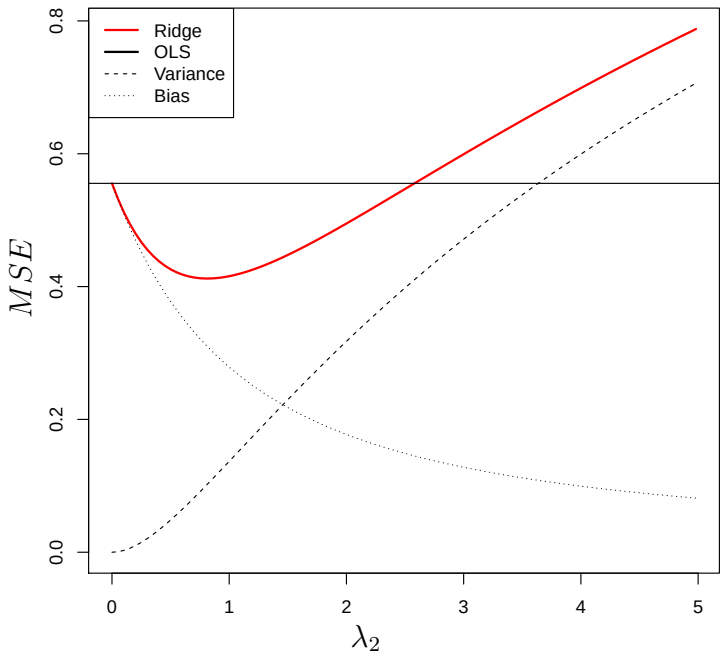


Figure 2.1 – Mean squared error for the ridge regression estimate according to the penalty parameter λ_2, and decomposition in (squared) bias and variance

where $\hat{\sigma}^2 = 1/(n-p)\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(0)\|^2$ is the classical unbiased estimation of the noise variance. Note that the consistency of ridge estimates requires λ_2 to converge towards zero when n goes to infinity.

These two estimates of λ_2 rely on the availability of the OLS regression estimates, which implies a $n > p$ setting. Some simple strategies like $\lambda_2 = 1/\sqrt{n}$ (Liu and Yu 2013) comply with the requirements for consistency, but may behave extremely poorly in high dimension. As a result, in many practical applications, λ_2 is set by simple cross-validation, although this effective heuristic does not come with consistency guaranties.

The bias of the ridge estimate is sometimes considered as a central issue when testing the nullity of the regression parameters. Some authors (Obenchein 1977, Bühlmann 2013) propose solutions to overcome this problem, but they do not address another important point, namely that the predictions $\mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_2)$ are not orthogonal projections of the responses, so that these predictions are not independent from the residuals.

2.3 Inaccuracy of Classical Tests

We advocate here that classical OLS tests are inaccurate for ridge regression when λ_2 differs from zero. Indeed, the properties of OLS and ridge statistics differ, because OLS essentially performs an orthogonal projection, leading to unbiased estimates. The shrinkage of ridge regression produces biased estimates and residuals that are correlated with the response. These well-known observations explain why the usual test statistics do not follow the well-known distributions for ridge estimates. Notice that in this section we consider $\mathbf{X}^\top\mathbf{X}$ as being invertible.

Student's Test While testing the nullity of β_j from the OLS estimates, we rely on an unbiased estimation of the parameter. Thus $\mathbb{E}(\hat{\beta}_j)$ is assumed to be zero when $\mathcal{H}_0$ holds. The Student statistic (1.4) is a ratio whose numerator is a random variable following a normal distribution centered in zero.

Concerning the denominator of the Student statistic, let us consider the OLS hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and the predicted response $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$:

$$\begin{aligned}
\mathbb{E}[\|\mathbf{y}-\hat{\mathbf{y}}\|^2] &= \operatorname{trace}(\mathbb{E}[(\mathbf{y}-\hat{\mathbf{y}})(\mathbf{y}-\hat{\mathbf{y}})^\top] \\
&= \operatorname{trace}((\mathbf{I}_n-\mathbf{H})\mathbb{E}[\mathbf{y}\mathbf{y}^\top](\mathbf{I}_n-\mathbf{H})^\top) \\
&= \operatorname{trace}((\mathbf{I}_n-\mathbf{H})(\mathbf{X}\boldsymbol{\beta}^\star\boldsymbol{\beta}^{\star\top}\mathbf{X}^\top+\sigma^2\mathbf{I}_n)(\mathbf{I}_n-\mathbf{H})^\top) \\
&= \sigma^2 \operatorname{trace}(\mathbf{I}_n-\mathbf{H}) \\
&= \sigma^2(n-p) \ ,
\end{aligned}$$

where we used the projection property $\mathbf{H}\mathbf{H} = \mathbf{H}$ and $\operatorname{trace}(\mathbf{H}) = \operatorname{trace}(\mathbf{I}_p) = p$. With the method of moments, we obtain

$$\hat{\sigma}^2 = \frac{1}{n-p}\|\mathbf{y}-\hat{\mathbf{y}}\|^2,$$

so the denominator of the Student statistic follows a χ^2 distribution with $n-p$ degrees of freedom. According Cochran's theorem, numerator

and denominator are independent:

$$\begin{aligned} cov[\hat{\mathbf{y}}, \mathbf{y} - \hat{\mathbf{y}}] =& cov[\mathbf{H}\mathbf{y}, (\mathbf{I}_n - \mathbf{H})\mathbf{y}] \\ =& \mathbf{H} cov[\mathbf{y}, \mathbf{y}](\mathbf{I}_n - \mathbf{H}) \\ =& \sigma^2 \mathbf{H}(\mathbf{I}_n - \mathbf{H}) \\ =& \sigma^2(\mathbf{H} - \mathbf{H}\mathbf{H}) \\ =& 0 \ . \end{aligned}$$

The covariance is null so $\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{y} - \hat{\mathbf{y}}$ and consequently $\hat{\boldsymbol{\beta}}(0) \perp\!\!\!\perp \mathbf{y} - \hat{\mathbf{y}}$ because $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}(0)$.

Halawa and El Bassiouni (1999) propose the following Student-like statistic for ridge regression:

$$\frac{\hat{\beta}_j(\lambda_2)}{S_j}$$

where S_j is the square root of the j^{th} diagonal element of the covariance matrix

$$\hat{\sigma}^2(\mathbf{X}^\top\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda_2\mathbf{I})^{-1} \ ,$$

with

$$\hat{\sigma}^2 = \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{n - \text{trace}(2\mathbf{H} - \mathbf{H}^2)} \ ,$$

where $\mathbf{H}$ is now the ridge hat matrix $\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}^\top$. Ridge estimates are biased so $\mathbb{E}[\hat{\beta}_j(\lambda_2)]$ may differs from zero when $\beta_j^* = 0$. Notice that estimating the bias is a difficult problem. As the hat matrix $\mathbf{H}$ is not a projection matrix, $\mathbf{H}^2 \neq \mathbf{H}$ and consequently

$$\begin{aligned} \mathbb{E}[(\mathbf{y} - \hat{\mathbf{y}})^\top(\mathbf{y} - \hat{\mathbf{y}})] =& \mathbb{E}[\mathbf{y}^\top(\mathbf{I}_n - \mathbf{H})^2\mathbf{y}] \\ =& \mathbb{E}[(\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon})^\top(\mathbf{I}_n - \mathbf{H})^2(\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon})^\top] \\ =& \sigma^2 \,\text{trace}((\mathbf{I}_n - \mathbf{H})^2) + (\mathbf{X}\boldsymbol{\beta}^*)^\top(\mathbf{I}_n - \mathbf{H})^2(\mathbf{X}\boldsymbol{\beta}^*) \ , \end{aligned}$$

without further simplifications. The distribution of the denominator is not a simple χ^2 distribution and is tricky to estimate.

We obtain also

$$cov[\hat{\mathbf{y}}, \mathbf{y} - \hat{\mathbf{y}}] = \sigma^2(\mathbf{H} - \mathbf{H}^2) \neq 0,$$

so $\hat{\mathbf{y}} \not\perp \mathbf{y} - \hat{\mathbf{y}}$ and $\hat{\boldsymbol{\beta}}(\lambda_2) \not\perp \mathbf{y} - \hat{\mathbf{y}}$. A random variable defined as a ratio of a normal to an independent χ^2 follows a student distribution. The dependency between the numerator and the denominator of the Student-like statistic is thus problematic: in ridge regression, especially when the penalty parameter is large, the considered statistic may be far from being Student distributed. The test may thus be inaccurate in terms of p-values as appears in the following sections. That being said, if λ_2 is sufficiently close to 0, then this test is acceptable as shown in the last section of this chapter.

Fisher's Test Using a Fisher test may be thought as a manner to alleviate the bias problem of the t-test. Indeed the numerator of the F-statistic (1.5) is a difference between the residuals of two models. If the bias is comparable in the two models, the bias in the difference will be small.

The Fisher distribution corresponds to the ratio of two χ^2 independent random variables. Unfortunately, for ridge regression, the numerator and denominator of the F-statistic are not independent and do not follow a χ^2 distribution. Again this test is unsuitable for ridge regression. However, the key principle of the Fisher's test, comparing two models, will be used for our proposed permutation test presented in the following section.

2.4 Resampling Approaches

In the general case, testing the hypothesis $\beta_j = 0$ from the ridge estimate is difficult since the distribution of each ridge coefficient depends on the whole vector of parameters $\boldsymbol{\beta}$. In such circumstances, resampling strategies are a convenient tool for approaching the distribution. This section presents different resampling strategies based on bootstrap or permutation. All methods focus on estimating the distribution of the classical Student's statistics or partial correlation coefficient, except for our proposal which considers Fisher's statistics

(see Table 2.1). Although both Student and Fisher statistics lead to equivalent tests in the OLS setting, this is no longer true when testing penalized coefficients.

Let us introduce a few notations, starting with the full model

$$\mathbf{y} = \mathbf{X}_{\boldsymbol{\omega}}\hat{\beta}_{\boldsymbol{\omega}} + \mathbf{X}_{\bar{\boldsymbol{\omega}}}\hat{\beta}_{\bar{\boldsymbol{\omega}}} + \boldsymbol{\epsilon} \tag{2.3}$$

where the full variable set $\Omega = \bar{\boldsymbol{\omega}} \cup \boldsymbol{\omega}$ is defined as the concatenation $\boldsymbol{\omega}$ the variable set defining the small model and $\bar{\boldsymbol{\omega}}$ set of remaining variables. The model which does not considers $\bar{\boldsymbol{\omega}}$ as explanatory variable is denoted as small or reduced model:

$$\mathbf{y} = \mathbf{X}_{\boldsymbol{\omega}}\hat{\beta}_{\boldsymbol{\omega}} + \boldsymbol{\epsilon}. \tag{2.4}$$

The question of finding wether the variables of $\bar{\boldsymbol{\omega}}$ are useful for explaining $\mathbf{y}$ can be formulated as a hypothesis test where the null hypothesis assumes $\boldsymbol{\beta}_{\bar{\boldsymbol{\omega}}}$ to be zero.

The large model (2.3) will be denoted $\mathbf{y}|\boldsymbol{\Omega}$. $\hat{\boldsymbol{\beta}}_{\mathbf{y}|\boldsymbol{\Omega}}$ corresponds to estimation of the parameter vector in the large model context and $\mathbf{S}_{\mathbf{y}|\boldsymbol{\Omega}}$ corresponds to $se(\hat{\boldsymbol{\beta}}_{\mathbf{y}|\boldsymbol{\Omega}})$, which denotes the standard deviation of the estimates. $\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}}$ denotes the prediction $\hat{\mathbf{y}} = \mathbf{X}_{\boldsymbol{\omega}}\hat{\boldsymbol{\beta}}(\lambda_2)_{\boldsymbol{\omega}} + \mathbf{X}_{\bar{\boldsymbol{\omega}}}\hat{\boldsymbol{\beta}}(\lambda_2)_{\bar{\boldsymbol{\omega}}}$ and $\mathbf{R}_{\mathbf{y}|\boldsymbol{\Omega}}$ denotes the corresponding residuals $\mathbf{y} - \hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}}$ for model $\mathbf{y}|\boldsymbol{\Omega}$. The reduce model will be denoted by same notations for $\mathbf{X}_{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\beta}}(\lambda_2)_{\boldsymbol{\omega}}$.

2.4.1 Bootstrapping

Regression models can be bootstrapped by considering the explanatory variables as random and selecting bootstrap samples directly from the observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ or treating the explanatory variables as fixed and resampling from the residuals of the fitted regression model.

Non-Parametric Bootstrap Crivelli et al. (1995) pick the first approach for computing confidence intervals. We thus denote this approach by $C95$. In this approach B bootstrap samples of size n are sampled with replacement from the original sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$. For

each bootstrapped sample b a bootstrapped estimates is computed:

$$\hat{\boldsymbol{\beta}}_{\mathbf{y}^{(b)}|\boldsymbol{\Omega}^{(b)}}. \tag{2.5}$$

In the following, we consider that the full model has only one additional variable compared to the reduced model: $\bar{\boldsymbol{\omega}} = \{j\}$. For the sake of simplicity we thus use interchangeably $\bar{\boldsymbol{\omega}}$ or j according the context. For each β_j and each bootstrap sample, a Studentized statistic is derived

$$t_j^{(b)} = \frac{\hat{\beta}_j^{(b)} - \hat{\beta}_j}{S_j^{(b)}}, \tag{2.6}$$

where $S_j^{(b)}$ is the classical standard deviation estimator computed from the bootstrapped sample. Let $\bar{\boldsymbol{\omega}}$ be a given quantile of the normal distribution. Using the B Studentized statistics Crivelli et al. (1995) compute the bootstrap probability to be smaller than $\bar{\boldsymbol{\omega}}$

$$\hat{P}(z) = \frac{\#\{t_j^{(b)} \leq z\}}{B},$$

and uses it to produce a confidence interval corrected according the approximation proposed by Rayner (1989):

$$[\hat{\beta}_j - u_j, \hat{\beta}_j + v_j],$$

where

$$\begin{cases} u_j = S_j z_{1-\alpha/2} - S_j \frac{(\hat{p}(z_{1-\alpha/2}) - (1-\alpha/2))}{\phi(z_{\alpha/2})}, \\ v_j = S_j z_{1-\alpha/2} + S_j \frac{(\hat{p}(z_{1-\alpha/2}) - \alpha/2)}{\phi(z_{\alpha/2})}. \end{cases}$$

Without relying on this approximation, which aims at correcting the distribution of the estimates, it is possible to take advantage of the bootstrap samples to derive a bootstrapped p-value for the test

$$\begin{cases} \mathcal{H}_0 : \beta_j^* = 0, \\ \mathcal{H}_1 : \beta_j^* \neq 0. \end{cases}$$

$$\begin{aligned} P(\hat{\beta}_j) &= \frac{2}{B} * min\left(\#\{\hat{\beta}_j < \hat{\beta}_j^b - \frac{\hat{\beta}_j}{S_j} S_j^b\}, \#\{\hat{\beta}_j > \hat{\beta}_j^b - \frac{\hat{\beta}_j}{S_j} S_j^b\}\right), \\ &= \frac{2}{B} * min\left(\#\{t_j^{(b)} > t_j\}, \#\{t_j^{(b)} < t_j\}\right). \end{aligned}$$

Residual Bootstrap Lopes (2014) studies the residual bootstrap for ridge regression in high-dimensional setting. This work is closely related to the idea presented by Chatterjee and Lahiri (2013) and it will be denoted by $L14$. The principle is the following

1. From the original sample compute the ridge estimate $\hat{\boldsymbol{\beta}}(\lambda_2)$
2. Compute the residuals $e_i = y_i - \hat{\boldsymbol{\beta}}(\lambda_2)^t \mathbf{x_i}$.
3. Consider the centered residuals $e_i - \bar{e}_i$.
4. Select a random sample $\{e_i^{(b)}\}_{i=1}^n$ of size n sampling the centered residuals with replacement.
5. Generate a sample $\{y_i^{(b)} = \hat{\boldsymbol{\beta}}(\lambda_2)^t \mathbf{x_i} + e_i^{(b)}\}_{i=1}^n$
6. Based on the bootstrap dataset $\{\mathbf{x_i}, y_i^{(b)}\}_{i=1}^n$ compute a bootstraped ridge estimate $\hat{\boldsymbol{\beta}}(\lambda_2)^{(b)}$.
7. Consider the bootstraped version $\mathbf{t}^{(b)} = \sqrt{n}(\hat{\boldsymbol{\beta}}(\lambda_2)^{(b)} - \hat{\boldsymbol{\beta}}(\lambda_2))$ of $\mathbf{t} = \sqrt{n}(\hat{\boldsymbol{\beta}}(\lambda_2) - \boldsymbol{\beta})$.

From all the B bootstraped statistics a p-value can be computed for each component j:

$$P(s\hat{\beta}_j) = \frac{2}{B} * min\left(\#\{t_j^{(b)} > t_j\}, \#\{t_j^{(b)} < t_j\}\right).$$

2.4.2 Permutation Strategies

Permutation is an alternative resampling method which has often been applied to OLS for testing coefficients (Anderson and Legendre 1999, Freedman and Lane 1983, Manly 2006, Kennedy 1995). Permutation is helpful when the sample does no fulfill assumptions required by traditional parametric testing procedures.

Permutation tests assume exchangeability under null hypothesis, which states that the joint probability distribution of the permuted samples is the same as the joint probability distribution of the original sample. Independent and identically-distributed random sample are exchangeable but samples with some simple dependance structure may also be exchangeable.

As for the bootstrap, permutation strategies consider different solutions according the quantities to be permuted, and also consider different test statistics. A taxonomy of permutation methods dedicated to regression coefficient testing can thus be achieved by considering the permutable unit and the test statistics.

Permutable unit is either the response $\mathbf{y}$, the predictor to test $\bar{\boldsymbol{\omega}}$ or a residual. An additional criterion for partitioning the available methods is the model which is used for permutation. Some author consider permutation in the full model $\mathbf{y}|\boldsymbol{\Omega}$, while other focus on the reduced model $\mathbf{y}|\mathbf{X}_{\boldsymbol{\omega}}$.

Test statistics classically used are the Student statistic, the Fisher statistic and squared partial correlation coefficient. Recall that in the context of OLS, considering Student or Fisher statistics is equivalent for testing a unique coefficient. Nevertheless this equivalence does not hold with ridge Regression.

Full Model Residual Permutations with T-Statistics The permutational analog of residual bootstrap uses the residuals from the full regression model $\mathbf{y}|\boldsymbol{\Omega}$ as permutation unit (Ter Braak 1992). New responses are calculated by adding the permuted residuals $\tilde{\mathbf{R}}^{(b)}_{\mathbf{y}|\boldsymbol{\Omega}}$ to the original estimation $\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}}$:

$$\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}} + \tilde{\mathbf{R}}^{(b)}_{\mathbf{y}|\boldsymbol{\Omega}}.$$

From these new samples, new parameters $\hat{\boldsymbol{\beta}}_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}}+\tilde{\mathbf{R}}^{(b)}_{\mathbf{y}|\boldsymbol{\Omega}}\right)|\boldsymbol{\Omega}}$ are estimated and the permutation based statistic

$$t^{(b)}_{\bar{\boldsymbol{\omega}}} = \frac{\hat{\boldsymbol{\beta}}_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}}+\tilde{\mathbf{R}}^{(b)}_{\mathbf{y}|\boldsymbol{\Omega}}\right)|\boldsymbol{\Omega}} - \hat{\boldsymbol{\beta}}_{\mathbf{y}|\boldsymbol{\Omega}}}{\mathbf{S}_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}}+\tilde{\mathbf{R}}^{(b)}_{\mathbf{y}|\boldsymbol{\Omega}}\right)|\boldsymbol{\Omega}}} \tag{2.7}$$

is computed. As in residual bootstrap the variability of the permutation based estimates $\hat{\boldsymbol{\beta}}_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}}+\tilde{\mathbf{R}}^{(b)}_{\mathbf{y}|\boldsymbol{\Omega}}\right)|\boldsymbol{\Omega}}$ around the original sample estimate $\hat{\boldsymbol{\beta}}_{\mathbf{y}|\boldsymbol{\Omega}}$ mimic the variability of $\hat{\boldsymbol{\beta}}_{\mathbf{y}|\boldsymbol{\Omega}}$ around the true value of the parameter. The distribution of permutation based statistic is then to

be compared to the sample statistic under $\mathcal{H}_0$:

$$t_{\bar{\boldsymbol{\omega}}} = \frac{\hat{\boldsymbol{\beta}}_{\mathbf{y}|\boldsymbol{\Omega}} - 0}{\mathbf{S}_{\hat{\boldsymbol{\beta}}_{\mathbf{y}|\boldsymbol{\Omega}}|}}. \tag{2.8}$$

This comparison allows to deduce a p-value (Anderson and Legendre 1999)

$$P_{\bar{\boldsymbol{\omega}}} = \frac{1}{B}\# \left\{ \left| \frac{\hat{\boldsymbol{\beta}}^{\bar{\boldsymbol{\omega}}}_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}} + \tilde{\mathbf{R}}^{(b)}_{\mathbf{y}|\boldsymbol{\Omega}}\right)|\boldsymbol{\Omega}} - \hat{\boldsymbol{\beta}}^{\bar{\boldsymbol{\omega}}}_{\mathbf{y}|\boldsymbol{\Omega}}}{\mathbf{S}^{\bar{\boldsymbol{\omega}}}_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}} + \tilde{\mathbf{R}}^{(b)}_{\mathbf{y}|\boldsymbol{\Omega}}\right)|\boldsymbol{\Omega}}} \right| \geq \left| \frac{\hat{\boldsymbol{\beta}}^{\bar{\boldsymbol{\omega}}}_{\mathbf{y}|\boldsymbol{\Omega}}}{\mathbf{S}^{\bar{\boldsymbol{\omega}}}_{\mathbf{y}|\boldsymbol{\Omega}}} \right| \right\}. \tag{2.9}$$

This approach will be denoted by $A99$.

Reduced Model Permutations with Partial Correlation Coefficient Instead of permuting under the full model it is possible to permute the residuals under the reduced model. The intuition for this permutation is that given some estimate of the relationship between $\mathbf{y}$ and $\mathbf{X}_{\boldsymbol{\omega}}$ there is no further variation in $\mathbf{y}$ which can be explained by $\mathbf{X}_{\bar{\boldsymbol{\omega}}}$.

Residual Permutations Kennedy (1995) explored permuting the residuals with the statistic

$$r^2 = \frac{\left(\sum \mathbf{R}_{\mathbf{y}|\mathbf{X}_{\boldsymbol{\omega}}} \mathbf{R}_{\mathbf{X}_{\bar{\boldsymbol{\omega}}}|\mathbf{X}_{\boldsymbol{\omega}}}\right)^2}{\sum \left(\mathbf{R}^2_{\mathbf{y}|\mathbf{X}_{\boldsymbol{\omega}}}\right) \sum \left(\mathbf{R}^2_{\mathbf{X}_{\bar{\boldsymbol{\omega}}}|\mathbf{X}_{\boldsymbol{\omega}}}\right)} \quad . \tag{2.10}$$

We thus denote this approach by K95. The authors propose a statistic where only the numerator changes:

$$r^{2(b)} = \frac{\left(\sum \mathbf{R}^{(b)}_{\mathbf{y}|\mathbf{X}_{\boldsymbol{\omega}}} \mathbf{R}_{\mathbf{X}_{\bar{\boldsymbol{\omega}}}|\mathbf{X}_{\boldsymbol{\omega}}}\right)^2}{\sum \left(\mathbf{R}^2_{\mathbf{y}|\mathbf{X}_{\boldsymbol{\omega}}}\right) \sum \left(\mathbf{R}^2_{\mathbf{X}_{\bar{\boldsymbol{\omega}}}|\mathbf{X}_{\boldsymbol{\omega}}}\right)}. \tag{2.11}$$

Residuals being exchangeable, we get a $\bar{\boldsymbol{\omega}}$ variable which is uncorrelated with $\mathbf{y}$ with a distribution under $\mathcal{H}_0$ Thus the pvalue for $\bar{\boldsymbol{\omega}}$ takes the following form:

$$\frac{1}{B}\# \left\{ \frac{\left(\sum \mathbf{R}^{(b)}_{\mathbf{y}|\mathbf{X}_{\boldsymbol{\omega}}} \mathbf{R}_{\mathbf{X}_{\bar{\boldsymbol{\omega}}}|\mathbf{X}_{\boldsymbol{\omega}}}\right)^2}{\sum \left(\mathbf{R}^2_{\mathbf{y}|\mathbf{X}_{\boldsymbol{\omega}}}\right) \sum \left(\mathbf{R}^2_{\mathbf{X}_{\bar{\boldsymbol{\omega}}}|\mathbf{X}_{\boldsymbol{\omega}}}\right)} \geq r^2 \right\} =$$

$$\frac{1}{B}\# \left\{ \left(\sum \mathbf{R}^{(b)}_{\mathbf{y}|\mathbf{X}_{\boldsymbol{\omega}}} \mathbf{R}_{\mathbf{X}_{\bar{\boldsymbol{\omega}}}|\mathbf{X}_{\boldsymbol{\omega}}}\right)^2 \geq \left(\sum \mathbf{R}_{\mathbf{y}|\mathbf{X}_{\boldsymbol{\omega}}} \mathbf{R}_{\mathbf{X}_{\bar{\boldsymbol{\omega}}}|\mathbf{X}_{\boldsymbol{\omega}}}\right)^2 \right\}. \tag{2.12}$$

Residual Permutations and Refitting Freedman and Lane (1983) propose a slightly more complex estimate where they use the residuals to generate new responses, which in turns are used to compute residuals. This approach will be denoted by $F83$:

$$r^{2(b)} = \frac{\left(\sum \mathbf{R}_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\mathbf{X}_\omega}+\mathbf{R}^{(b)}_{\mathbf{y}|\mathbf{X}_\omega}\right)|\mathbf{X}_\omega}\mathbf{R}_{\mathbf{X}_{\bar{\omega}}|\mathbf{X}_\omega}\right)^2}{\sum\left(\mathbf{R}^2_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\mathbf{X}_\omega}+\mathbf{R}^{(b)}_{\mathbf{y}|\mathbf{X}_\omega}\right)|\mathbf{X}_\omega}\right)\sum\left(\mathbf{R}^2_{\mathbf{X}_{\bar{\omega}}|\mathbf{X}_\omega}\right)}. \tag{2.13}$$

This approach is close to the proposal of Kennedy (1995) if we assume that $\mathbf{R}^{(b)}_{\mathbf{y}|\mathbf{X}_\omega} \simeq \mathbf{R}^{(b)}_{\mathbf{y}|\mathbf{X}_\omega}\mathbf{R}_{\mathbf{X}_{\bar{\omega}}|\mathbf{X}_\omega}$. Though Anderson and Robinson (2001) showed by simulation in the context of OLS that the smaller n the larger the difference between the approaches. If the statistics K95 is much simpler to compute, it seems less accurate compared to the exact test F83 (Anderson and Robinson 2001).

This approach provides also a p-value for each subset $\bar{\boldsymbol{\omega}}$

$$\frac{1}{B}\#\left\{\frac{\left(\sum \mathbf{R}_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\mathbf{X}_\omega}+\mathbf{R}^{(b)}_{\mathbf{y}|\mathbf{X}_\omega}\right)|\mathbf{X}_\omega}\mathbf{R}_{\mathbf{X}_{\bar{\omega}}|\mathbf{X}_\omega}\right)^2}{\sum\left(\mathbf{R}^2_{\left(\hat{\mathbf{y}}_{\mathbf{y}|\mathbf{X}_\omega}+\mathbf{R}^{(b)}_{\mathbf{y}|\mathbf{X}_\omega}\right)|\mathbf{X}_\omega}\right)\sum\left(\mathbf{R}^2_{\mathbf{X}_{\bar{\omega}}|\mathbf{X}_\omega}\right)} \geq r^2\right\}. \tag{2.14}$$

Reduced Model and Permutations of Response Manly (2006) proposed another intuitive approach, which consists in permuting the $\mathbf{y}$ in the reduced model to get residuals $\mathbf{R}_{\mathbf{y}^{(b)}|\mathbf{X}_\omega}$ independent of residuals $\mathbf{R}_{\mathbf{X}_{\bar{\omega}}|\mathbf{X}_\omega}$. This approach will be denoted by $M06$. The statistics resulting from this permutation can be written as

$$r^{2(b)} = \frac{\left(\sum \mathbf{R}_{\mathbf{y}^{(b)}|\mathbf{X}_\omega}\mathbf{R}_{\mathbf{X}_{\bar{\omega}}|\mathbf{X}_\omega}\right)^2}{\sum\left(\mathbf{R}^2_{\mathbf{y}^{(b)}|\mathbf{X}_\omega}\right)\sum\left(\mathbf{R}^2_{\mathbf{X}_{\bar{\omega}}|\mathbf{X}_\omega}\right)}. \tag{2.15}$$

Notice that this last approach assumes that the $\mathbf{y}$ are exchangeable, which is obviously wrong (Anderson and Robinson 2001).

Predictor Permutations and F-Test Although it is widely used for model selection in penalized regression problems (for calibration and degrees of freedom issues, see Hastie and Tibshirani 1990), the F-test is not exact for ridge regression, for the reasons already mentioned

above (estimation bias and dependency between $(\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}} - \hat{\mathbf{y}}_{\mathbf{y}|\mathbf{x}_{\boldsymbol{\omega}}})$ and $(\mathbf{y} - \hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}})$). This approach will be denoted by $B15$. Here, we propose to approach the distribution of the derived F-statistic under the null hypothesis, which is defined as follow:

$$F_{\bar{\boldsymbol{\omega}}} = \frac{\|\hat{\mathbf{y}}_{\mathbf{y}|\mathbf{x}_{\boldsymbol{\omega}}} - \mathbf{y}\|^2 - \|\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}} - \mathbf{y}\|^2}{\|\hat{\mathbf{y}}_{\mathbf{y}|\boldsymbol{\Omega}} \ - \mathbf{y}|^2}. \tag{2.16}$$

Notice that degrees of freedom not appear as in the original F-statistic. To approach this distribution a randomization is computed as follow:

$$F^{(b)}_{\bar{\boldsymbol{\omega}}} = \frac{\|\hat{\mathbf{y}}_{\mathbf{y}|\mathbf{x}_{\boldsymbol{\omega}}} - \mathbf{y}\|^2 - \|\hat{\mathbf{y}}_{\mathbf{y}|\mathbf{x}_{\boldsymbol{\omega}}\mathbf{x}^{(b)}_{\bar{\boldsymbol{\omega}}}} - \mathbf{y}\|^2}{\|\hat{\mathbf{y}}_{\mathbf{y}|\mathbf{x}_{\boldsymbol{\omega}}\mathbf{x}^{(b)}_{\bar{\boldsymbol{\omega}}}} \ - \mathbf{y}|^2}. \tag{2.17}$$

Thereby, the permutation F-test, denoted B15 as follows, provides a p-value for each subset $\bar{\boldsymbol{\omega}}$

$$P_{\bar{\boldsymbol{\omega}}} = \frac{1}{B}\# \left\{ \frac{\left(\sum \mathbf{R}_{\mathbf{y}|\mathbf{x}_{\boldsymbol{\omega}}}\right)^2 - \left(\sum \mathbf{R}_{\mathbf{y}|\mathbf{x}_{\boldsymbol{\omega}}\mathbf{x}^{(b)}_{\bar{\boldsymbol{\omega}}}}\right)^2}{\left(\sum \mathbf{R}_{\mathbf{y}|\mathbf{x}_{\boldsymbol{\omega}}\mathbf{x}^{(b)}_{\bar{\boldsymbol{\omega}}}}\right)^2} \geq \frac{\left(\sum \mathbf{R}_{\mathbf{y}|\mathbf{x}_{\boldsymbol{\omega}}} - \mathbf{R}_{\mathbf{y}|\boldsymbol{\Omega}}\right)^2}{\left(\sum \mathbf{R}_{\mathbf{y}|\boldsymbol{\Omega}}\right)^2} \right\} \tag{2.18}$$

Permutation tests are often used in a small sample setting where Gaussian approximations of the maximum likelihood estimates are not valid. To be exact, permutation tests assume some form of exchangeability. There is no finite-sample exact permutation test in multiple linear regression (Anderson and Robinson 2001). A test based on partial residuals (under the null hypothesis regression model) is asymptotically exact for unpenalized regression, but it does not apply to penalized regression. Instead, we directly permute the values taken by the explicative variable to be tested, so as to estimate the distribution of the F-statistic under the null hypothesis that the variable is irrelevant. This permutation test is asymptotically exact when the tested variable is independent from the other explicative variables, and is approximate in the general case.

Table 3.3 shows that, compared to the standard t-test and F-test (Hastie and Tibshirani 1990), the permutation test provides a satisfactory control of the significance level. It is either well-calibrated or slightly more conservative than the prescribed significance level,

whereas the standard t-test and F-test result in false positive rates that are way above the asserted significance level, especially for strong correlations between explanatory variables. These observations apply throughout the experiments reported in Section 2.6.1. Note that testing all variables results in a multiple testing problem. We propose here to control the false discovery rate (FDR), which is defined as the expected proportion of false discoveries among all discoveries. This control requires to correct the p-values for multiple testing (Benjamini and Hochberg 1995). The overall procedure is well calibrated as shown in Section 2.6.

Permutation tests rely on the fitting of several hundredth of randomized models. The following section details our efficient implementation that drastically reduces the computational cost.

Table 2.1 – Summary of resampling strategies, classified according to total and partial correlation based statistics. The second column shows how resampling affects the model. The third column states the method associated to this strategy. Bootstrap and permutations strategies are tagged by a star.

Correlation	Resampling Model	Name
Total	$\mathbf{y}\|\mathbf{X}_\omega\mathbf{X}_{\bar{\omega}}^{(b)}$	B15
	$\mathbf{y}^{(b)}\|\mathbf{X}_\omega^{(b)}\mathbf{X}_{\bar{\omega}}^{(b)}$	C95*
	$\left(\hat{\mathbf{y}}_{\mathbf{y}\|\mathbf{\Omega}} + \tilde{\mathbf{R}}_{\mathbf{y}\|\mathbf{\Omega}}^{(b)}\right)\|\mathbf{\Omega}$	A99, L14*
Partial	$\left(\hat{\mathbf{y}}_{\mathbf{y}\|\mathbf{X}_\omega} + \mathbf{R}_{\mathbf{y}\|\mathbf{X}_\omega}^{(b)}\right)\|\mathbf{X}_\omega$	F83
	$\mathbf{y}^{(b)}\|\mathbf{X}_\omega$	M06
	$\mathbf{R}_{\mathbf{y}\|\mathbf{X}_\omega}^{(b)}$	K95

2.5 Efficient Implementation

2.5.1 Computations

Permutation tests rely on the simulation of numerous data sampled under the null hypothesis distribution. The number of replications must be important to estimate the rather extreme quantiles we are typically interested in. Here, we use $B = 1000$ replications for each variable or subset of variables to test. Each replication involves the fitting of a model where the complexity is in $O(p^2n + p^3)$, corresponding to the computation and inversion of $\mathbf{X}^\top\mathbf{X}$ respectively. When $n > p$ the total complexity is equal to p^2n because p^3 is dominated. But in the high-dimension, p^3 becomes dominant.

If we test each variable separately, the total computational cost for solving these B systems of size p with p variables is $O(Bp(p^3 + p^2n))$. Solving systems with a Cholesky decomposition of $\mathbf{X}^\top\mathbf{X}$ yields a complexity of $O(Bp(\frac{p^3}{3} + p^2n))$ (DO Q 2012).However, computation time can be saved using a block-wise decomposition and inversion.

The ridge estimate is computed as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\mathbf{X}^\top\mathbf{y} \ ,$$

where $\boldsymbol{\Lambda}$ is the diagonal penalty matrix whose jth diagonal entry is λ_2.

In the F-statistic (1.5), the permutation affects the calculation of the larger model residual RSS^Ω, which is denoted $\text{RSS}^{\Omega(b)}$ for the bth permutation. Using a similar notation for the design matrix and the estimated parameters, we have $\text{RSS}^{\Omega(b)} = \left\|\mathbf{y} - \mathbf{X}^{(b)}\hat{\boldsymbol{\beta}}^{(b)}\right\|_2^2$. $\mathbf{X}^{(b)}$ is defined as the concatenation of the rows permuted matrix $\mathbf{X}^{(b)}_{\bar{\boldsymbol{\omega}}}$ and the other original variables $\mathbf{X}_{.\boldsymbol{\omega}}$, where $\boldsymbol{\omega} = \{j : j \notin \bar{\boldsymbol{\omega}}\}$: $\mathbf{X}^{(b)} = (\mathbf{X}^{(b)}_{\bar{\boldsymbol{\omega}}}, \mathbf{X}_{.\boldsymbol{\omega}})$. Then, $\hat{\boldsymbol{\beta}}^{(b)}$ can be efficiently computed by using $\underset{|\boldsymbol{\omega}|\times|\boldsymbol{\omega}|}{\mathbf{D}} \in \mathbb{R}^{|\boldsymbol{\omega}|}$,$\underset{|\bar{\boldsymbol{\omega}}|\times|\bar{\boldsymbol{\omega}}|}{\mathbf{A}} \in \mathbb{R}^{|\bar{\boldsymbol{\omega}}|}$ and $\underset{|\boldsymbol{\omega}|\times|\bar{\boldsymbol{\omega}}|}{\mathbf{U}} \in \mathbb{R}^{|\bar{\boldsymbol{\omega}}|}$ defined as follows:

$$\begin{aligned}
\underset{|\boldsymbol{\omega}|\times|\boldsymbol{\omega}|}{\mathbf{D}} &= {\mathbf{X}_{.\boldsymbol{\omega}}}^\top\mathbf{X}_{.\boldsymbol{\omega}} + \boldsymbol{\Lambda}_{\boldsymbol{\omega}\boldsymbol{\omega}}, \\
\underset{|\bar{\boldsymbol{\omega}}|\times|\bar{\boldsymbol{\omega}}|}{\mathbf{A}} &= {\mathbf{X}^{(b)}_{.\bar{\boldsymbol{\omega}}}}^\top\mathbf{X}^{(b)}_{.\bar{\boldsymbol{\omega}}} + \boldsymbol{\Lambda}_{\bar{\boldsymbol{\omega}}\bar{\boldsymbol{\omega}}} - {\mathbf{X}^{(b)}_{.\bar{\boldsymbol{\omega}}}}^\top\mathbf{X}_{.\boldsymbol{\omega}}\mathbf{D}^{-1}\mathbf{X}^\top_{.\boldsymbol{\omega}}\mathbf{X}^{(b)}_{.\bar{\boldsymbol{\omega}}} \text{ and} \\
\underset{|\boldsymbol{\omega}|\times|\bar{\boldsymbol{\omega}}|}{\mathbf{U}} &= -\mathbf{D}^{-1}\mathbf{X}^\top_{.\boldsymbol{\omega}}\mathbf{X}^{(b)}_{.\bar{\boldsymbol{\omega}}} \ .
\end{aligned}$$

Indeed, using the Schur complement, one writes $\hat{\boldsymbol{\beta}}^{(b)}$ as follows:

$$\hat{\boldsymbol{\beta}}^{(b)} = \left(\begin{pmatrix} \mathbf{I}_{|\bar{\boldsymbol{\omega}}|} & \mathbf{0}_{|\bar{\boldsymbol{\omega}}|\times|\boldsymbol{\omega}|} \\ \mathbf{U} & \mathbf{I}_{|\boldsymbol{\omega}|} \end{pmatrix} \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{0}_{|\bar{\boldsymbol{\omega}}|\times|\boldsymbol{\omega}|} \\ \mathbf{0}_{|\boldsymbol{\omega}|\times|\bar{\boldsymbol{\omega}}|} & \mathbf{D}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{|\boldsymbol{\omega}|} & \mathbf{U}^\top \\ \mathbf{0}_{|\boldsymbol{\omega}|\times|\bar{\boldsymbol{\omega}}|} & \mathbf{I}_{|\bar{\boldsymbol{\omega}}|} \end{pmatrix} \right) \begin{pmatrix} {\mathbf{X}^{(b)}_{\cdot\bar{\boldsymbol{\omega}}}}^\top \mathbf{y} \\ {\mathbf{X}_{\cdot\boldsymbol{\omega}}}^\top \mathbf{y} \end{pmatrix}$$

$$= \left(\begin{pmatrix} \mathbf{I}_{|\bar{\boldsymbol{\omega}}|} \\ \mathbf{U} \end{pmatrix} \mathbf{A}^{-1} \begin{pmatrix} \mathbf{I}_{|\bar{\boldsymbol{\omega}}|} \\ \mathbf{U} \end{pmatrix}^\top + \begin{pmatrix} \mathbf{0}_{|\bar{\boldsymbol{\omega}}|\times|\bar{\boldsymbol{\omega}}|} & \mathbf{0}_{|\bar{\boldsymbol{\omega}}|\times|\boldsymbol{\omega}|} \\ \mathbf{0}_{|\boldsymbol{\omega}|\times|\bar{\boldsymbol{\omega}}|} & \mathbf{D}^{-1} \end{pmatrix} \right) \begin{pmatrix} {\mathbf{X}^{(b)}_{\cdot\bar{\boldsymbol{\omega}}}}^\top \mathbf{y} \\ {\mathbf{X}_{\cdot\boldsymbol{\omega}}}^\top \mathbf{y} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{I}_{|\bar{\boldsymbol{\omega}}|} \\ \mathbf{U} \end{pmatrix} \mathbf{A}^{-1} \begin{pmatrix} \mathbf{I}_{|\bar{\boldsymbol{\omega}}|} \\ \mathbf{U} \end{pmatrix}^\top \begin{pmatrix} {\mathbf{X}^{(b)}_{\cdot\bar{\boldsymbol{\omega}}}}^\top \mathbf{y} \\ {\mathbf{X}_{\cdot\boldsymbol{\omega}}}^\top \mathbf{y} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{|\bar{\boldsymbol{\omega}}|\times 1} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{\omega}} \end{pmatrix} \quad .$$

Hence, $\hat{\boldsymbol{\beta}}^{(b)}$ can be obtained as a correction of the vector of coefficients $\hat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}$ obtained under the smaller model. The key observation to be made here is that $\mathbf{X}_{\cdot\bar{\boldsymbol{\omega}}}$ does not intervene in the expression $\mathbf{D}^{-1}$, which is the bottleneck in the computation of $\mathbf{A}$, $\mathbf{U}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\omega}} = \mathbf{D}^{-1}{\mathbf{X}_{\cdot\boldsymbol{\omega}}}^\top \mathbf{y}$. It can therefore be performed once for all permutations. Additionally, $\mathbf{D}^{-1}$ can be cheaply computed from $\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}$ as follows:

$$\left({\mathbf{X}_{\cdot\boldsymbol{\omega}}}^\top \mathbf{X}_{\cdot\boldsymbol{\omega}} + \boldsymbol{\Lambda}_{\boldsymbol{\omega}\boldsymbol{\omega}}\right)^{-1} = \left[\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{\boldsymbol{\omega}\boldsymbol{\omega}} - \left[\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{\boldsymbol{\omega}\bar{\boldsymbol{\omega}}} \left[\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]^{-1}_{\bar{\boldsymbol{\omega}}\bar{\boldsymbol{\omega}}} \left[\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}\right]_{\bar{\boldsymbol{\omega}}\boldsymbol{\omega}} \quad .$$

Thus we compute $\left(\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Lambda}\right)^{-1}$ once, firstly correct for the removal of variables in $\bar{\boldsymbol{\omega}}$, secondly correct for permutation b, thus eventually requiring $O(B(p^2 + (p - |\bar{\boldsymbol{\omega}}|)n)))$ operations, to test one subset $\bar{\boldsymbol{\omega}}$, as it summarize in the table 2.2. We need 1000 permutation to have a precision at the percent order.

Table 2.2 – Breakdown of complexity for the efficient and vanilla approaches to compute the ridge estimate on B permutations for a single group of size $|\bar{\boldsymbol{\omega}}|$.

Approach	Step	Complexity	Occurrence
Vanilla	Compute $\hat{\boldsymbol{\beta}}^{(b)}$	$O(\frac{p^3}{3} + p^2 n)$	B
Efficient	Solve system	$O(p^3 + p^2 n)$	1
	Remove $\bar{\boldsymbol{\omega}}$	$O((p - \lvert\bar{\boldsymbol{\omega}}\rvert)^2)$	1
	Compute $\hat{\boldsymbol{\beta}}^{(b)}$	$O(p^2 + (p - \lvert\bar{\boldsymbol{\omega}}\rvert)n)$	B

To simplify the writing of complexity, on the overall process, we can consider $p - |\bar{\boldsymbol{\omega}}| = p$ where the size of the group of tested variables is ignored. That induce an overestimate of the real complexity to compute $\hat{\boldsymbol{\beta}}^{(b)}$ which becomes $O(p^2+pn)$. Finally, the global complexity becomes $O\left(p^3 + p^2(n + BG) + pnBG\right)$ where G is the number of tested groups and the vanilla approach has complexity $O(p^3\frac{BG}{3} + p^2nBG)$.

.

Table 2.3 – Breakdown of complexity for the efficient and vanilla approaches to compute the ridge estimate on B permutations for p variables tested independently. The last column summarizes the overall complexity.

Approach	Step	Complexity	Occ.	Total
Vanilla	Compute $\hat{\boldsymbol{\beta}}^{(b)}$	$O(\frac{p^3}{3} + p^2n)$	$B \times p$	$O\left(B(\frac{p^4}{3} + p^3n)\right)$
Efficient				
Optimized	Solve system	$O(p^3 + p^2n)$	1	$O\left(B(p^3 + p^2n)\right)$
	Remove $\bar{\boldsymbol{\omega}}$	$O(p^2)$	p	
	Compute $\hat{\boldsymbol{\beta}}^{(b)}$	$O(p^2 + pn)$	$B \times p$	

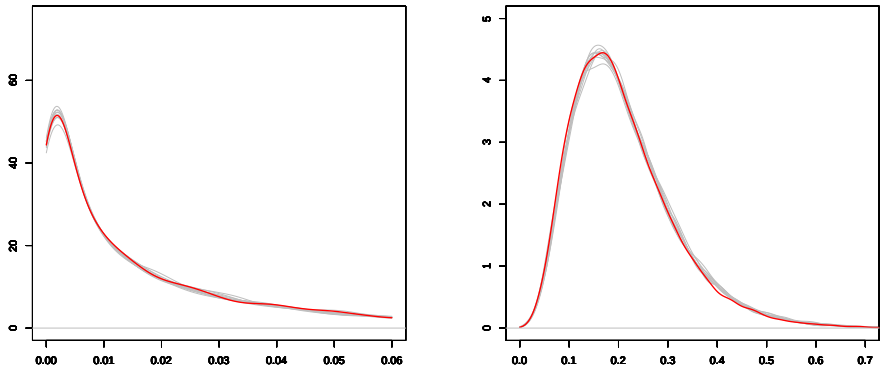
When each variable is tested independently, this approach provides an overall gain of factor p compared to the vanilla version with $O\left(B(p^3 + p^2n)\right)$ instead of $O\left(B(\frac{p^4}{3} + p^3n)\right)$ (see details in Table 2.3).

2.5.2 Homogeneity of Distributions Obtained by Permutations

Let us recall that, the complexity with our optimized algorithm for permutations is $O((Bq + p)(p^2 + pn))$ for q groups of variables tested with B permutations for each. In the following, we consider G groups with equal size $|\bar{\boldsymbol{\omega}}|$ to simplify the explanation.

If $n > p$ and $\lambda_2 = 0$ then the distribution of $\hat{F}_{(b)}$ along all permutations is like as an random variable $F\frac{|\bar{\boldsymbol{\omega}}|}{n-p}$ where $F \sim \mathcal{F}(|\bar{\boldsymbol{\omega}}|, n-p)$. That corresponds to permutations on OLS estimates where the exact-statistic F is scaled by $\frac{|\bar{\boldsymbol{\omega}}|}{n-p}$, as it shown in the figure 2.2. Considering $\mathbf{X}$ following a normal distribution, all group permutations follow the

same distribution and it will be tempting to learn a unique distribution for all groups.



$|\bar{\boldsymbol{\omega}}| = 1$ $|\bar{\boldsymbol{\omega}}| = 10$

Figure 2.2 – Density of statistics $\hat{F}_{(b)}$ obtained by permutations (grey lines) and of statistics $F\dfrac{|\bar{\boldsymbol{\omega}}|}{50}$ (red line) where $F \sim \mathcal{F}(|\bar{\boldsymbol{\omega}}|, 50)$ distribution. Density curves are computed for two groups size $|\bar{\boldsymbol{\omega}}|$, respectively 1, at left, and 10 at right. All permutations results are computed with $n = 250$, $p = 500$ and $B = 10.000$ for 10 different groups.

In the ortonormal case, ridge estimates are equal to OLS estimates with a factor dependent on λ_2 where

$$\hat{\boldsymbol{\beta}}(\lambda_2) = \frac{\hat{\boldsymbol{\beta}}(0)}{1+\lambda_2}.$$

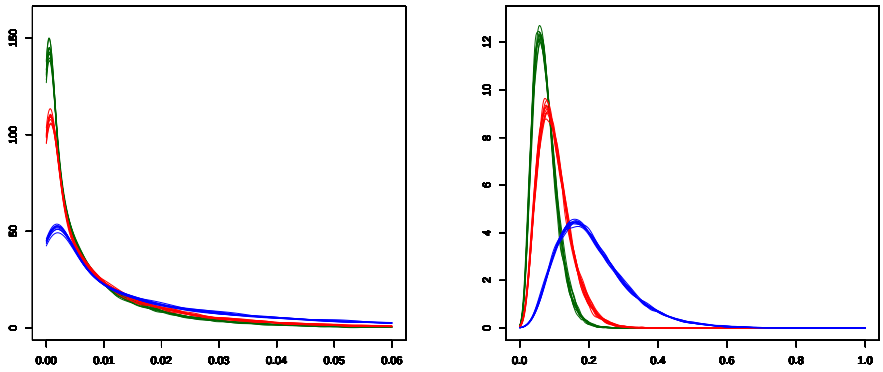
Then the ridge hat matrix can be rewritten as

$$\mathbf{H}_{\lambda_2} = \frac{\mathbf{H}_0}{1+\lambda_2}$$

and the residuals sum of square as

$$\mathrm{RSS}_{\lambda_2} = \mathbf{y}^\top \left(I - \frac{\mathbf{H}_0(1+2\lambda_2)}{(1+{\lambda_2}^2)} \right) \mathbf{y},$$

where $\mathbf{H}_0$ is the OLS hat matrix. Moreover, the classical F-test is constructed on the RSS and degrees of freedom are learned on the hat matrix. If we consider residuals of ridge regression with a close comportment than residuals of OLS, especially for the ortonormal case, then we can expect a close distribution of $\hat{F}_{(b)}$ under the null hypothesis



$|\bar{\boldsymbol{\omega}}| = 1$ $|\bar{\boldsymbol{\omega}}| = 10$

Figure 2.3 – Density of statistics $\hat{F}_{(b)}$ obtained by permutations. Density curves are computed for two groups size $|\bar{\boldsymbol{\omega}}|$, respectively 1, at left, and 10 at right. These are also computed for three level of penalties where $\lambda_2 = 0$ (blue lines), $\lambda_2 = 10$ (green lines) and $\lambda_2 = 20$ (red line). All results are computed with $n = 250$, $p = 500$ and $B = 10.000$ for 10 different simulations.

for each groups with same size $|\bar{\boldsymbol{\omega}}|$ and for a same penalty λ_2 even if we do not know its parameters and its shape.

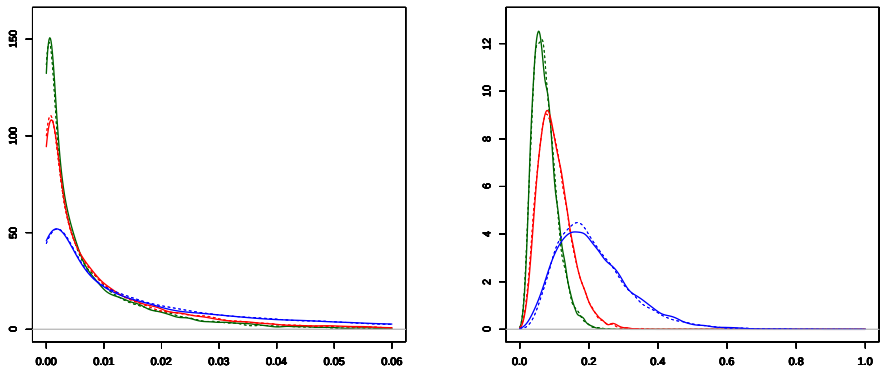
Figure 2.3 shows $\hat{F}_{(b)}$ distributions for different levels of penalties and different sizes $|\bar{\boldsymbol{\omega}}|$. There is a specific distribution for each couple penalty/size, so one distribution seems sufficient to estimate the comportment of $F_{(b)}$ under the null hypothesis for a group size and a specific penalty. If each variables are tested independently the overall complexity $O(Bp^3 + Bp^2n)$ became $O(p^3 + p^2(B + n) + Bpn)$ with almost one factor p is won.

Of course this approximation is only applicable for a constant penalty, so that will can not be applied on the adaptive-ridge presented on chapters 3 and 4.

2.5.3 Gamma Law Approximation

In the following subsection, we propose to do an empirical approximation of the $p-$values distribution by a gamma law, to do less permutation for a same estimation of the distribution of our statistic under

the null hypothesis.



$|\bar{\boldsymbol{\omega}}| = 1$ $|\bar{\boldsymbol{\omega}}| = 10$

Figure 2.4 – Density of statistics $\hat{F}_{(b)}$ obtained by permutations (full lines) and of values from Γ distribution learned on permutations (dashed lines). Density curves are computed for two groups size $|\bar{\boldsymbol{\omega}}|$, respectively 1, at left, and 10 at right. These are also computed for three level of penalties where $\lambda_2 = 0$(blue lines), $\lambda_2 = 10$ (green lines) and $\lambda_2 = 20$ (red line). All results are computed with $n = 250$, $p = 500$ and $B = 10.000$, only one simulation is shown.

$F-$statistics under the null hypothesis computed by permutation on ridge estimates not follows any Fisher distribution, as which was explain previously, even if the distribution of these statistics under permutations seems close to a Fisher distribution. Due to the capacity of the gamma law to have multiple shape, we look if the gamma distribution could be a good estimation of our statistic distribution under the null hypothesis. The figure 2.4 shows the distribution of our statistic with 10.000 permutations and the distribution of the gamma distribution where its parameter was learned on our statistic with 10.000 permutations. Parameters of the gamma distribution was obtained by the method of moments where

$$k = \frac{\overline{F}^2}{\overline{F^2} - \overline{F}^2}$$

and

$$\theta = \frac{\overline{F^2} - \overline{F}^2}{-\overline{F}}.$$

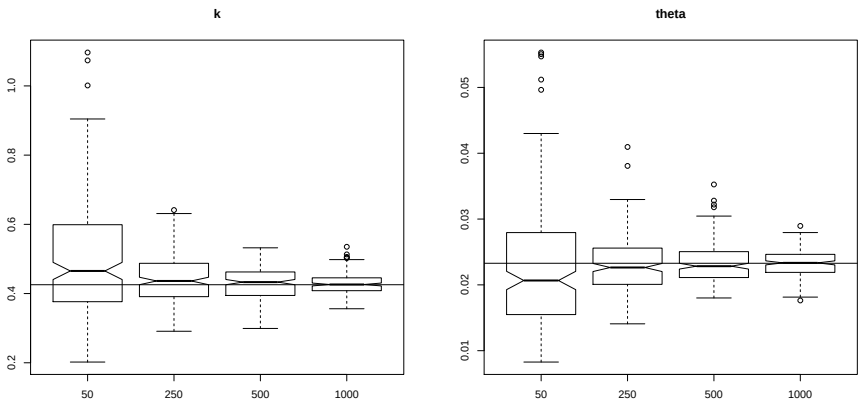


Figure 2.5 – Boxplots of distributions of shape parameter k (left side) and scale parameter θ (right side) for a gamma law approximation by the method of moments on $F_{(b)}$. Each boxplot is learned on $200 simulations$ with 50,250,500 and 1000 points used to learn the law. The horizontal line represents these parameters estimated on 100.000 permutations.

We can see an overlaping of curves which reassure us to use the strong assumption that our distribution of statistics under the null hypothesis is close to a gamma distribution. We took the gamma distribution instead a Fisher distribution because our statistic corresponds to a F-statistic scaled by $\frac{\text{df1}}{\text{df2}}$.

Using the gamma law approximation, to estimate the F distribution under the null hypothesis, needs to know the minimal number of necessary points to have a good estimation of k and θ by the method of moments. Figure 2.5 shows the variability of estimated parameters for different numbers of used points to learn gamma distribution and 250 or 500 permutations seem sufficient to learn the twice parameters. Indeed, expected value of k and θ, learned on 100.000 permutations, are very close to the mediane with a weak variability.

In fact, parameters of the hypothetic gamma law are obvious but we need to reject the null hypothesis as a same manner regardless of permutations or gamma distribution. Table 2.4 shows the number of rejected test based on permutations or gamma distributions and a gamma distribution based on 250 permutations seems a good alterna-

tive compared to a distribution based on 1000 permutations. In conclusion, we can use a gamma approximation to do test with a same level of confidence and that can be conjugate this gain with optimization already presented.

.

Table 2.4 – $\overline{|\mathcal{S}|}$ average number of variables which had a $p-$values <= 5% and $S(|\mathcal{S}|)$ its standard deviation. Result for gamma distribution estimated for different value of B and for $p-$values calculated directly on permutation. Obtained on 200 simulations for non 100000 distribution.

		50	250	500	1000	100000
Gamma	$\overline{\lvert\mathcal{S}\rvert}$	26.720	26.000	25.885	25.960	26.000
	$S(\lvert\mathcal{S}\rvert)$	1.047	0.549	0.415	0.242	—
Permutation	$\overline{\lvert\mathcal{S}\rvert}$	26.760	25.905	25.875	25.920	26.000
	$S(\lvert\mathcal{S}\rvert)$	1.144	0.654	0.501	0.307	—

2.6 Simulation of High Dimensionnal Effect

Previously, different approaches based on classical test, model resampling or residual resampling have been presented. In this section, we compare these approaches applied on the ridge regression with a focus on:

— The type I error and its effective control,

— The type II error.

2.6.1 Simulated Data

We consider the linear regression model $Y = X\boldsymbol{\beta}^\star + \varepsilon$, where Y is a continuous response variable, $X = (X_1, \ldots, X_p)^T$ is a vector of p predictor variables, $\boldsymbol{\beta}^\star$ is the vector of unknown parameters and ε is a zero-mean Gaussian error variable with variance σ^2. The parameter $\boldsymbol{\beta}^\star$ is sparse, that is, the support set $\mathcal{S}^\star = \left\{j \in \{1, ..., p\} | \beta_j^\star \neq 0\right\}$ indexing its non-zero coefficients is small $|\mathcal{S}^\star| \ll p$.

Simulation Models. Variable selection is known to be affected by numerous factors: the number of examples n, the number of variables p, the sparseness of the model $|\mathcal{S}^\star|$, the correlation structure of the explicative variables, the signal-to-noise ratio (SNR), the relative magnitude of the relevant parameters $\{\beta_j^\star\}_{j\in\mathcal{S}^\star}$, and the structure of the design matrix.

In our experiments, we defined $p = 100$, $|\mathcal{S}^\star| = 10$, $\rho = 0.5$ and we varied $n \in \{10, 100, 200\}$.

We considered four predictor correlation structures (schematized in the figure 2.6):

IND independent explicative variables following a zero-mean, unit-variance Gaussian distribution: $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

BLOCK dependent explicative variables following a zero-mean Gaussian distribution, with a block-diagonal covariance matrix: $X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\Sigma_{ii} = 1$, $\Sigma_{ij} = \rho$ for all pairs (i, j), $j \neq i$ belonging to the same block and $\Sigma_{ij} = 0$ for all pairs (i, j) belonging to different blocks. Each block is composed of 10 variables.

The position of relevant variables is dissociated from the block structure, that is, randomly distributed in $\{1, ..., p\}$. This design is thus difficult for variable selection.

GROUP same as BLOCK, except that the relevant variables are gathered in a single block, thus facilitating group variable selection.

TOEP$^-$ same as GROUP, except that $\Sigma_{ij} = -\rho^{|i-j|}$ for all pairs (i, j), $j \neq i$ belonging to the same block and $\Sigma_{ij} = 0$ for all pairs (i, j) belonging to different blocks.

We fixed the signal-to-noise ratio $\boldsymbol{\beta}^{\star\top}\mathbf{\Sigma}\boldsymbol{\beta}^\star/\sigma^2 = 4$, which leads to feasible but challenging problems for model selection. Finally, the non-zero parameters $\beta_j^\star$ are drawn from a uniform distribution $\mathcal{U}(10^{-1}, 1)$ to enable different magnitudes.

In the following we discuss only testing variables and not of groups

of variables. Indeed, the group question will be discuss in the chapter 4.

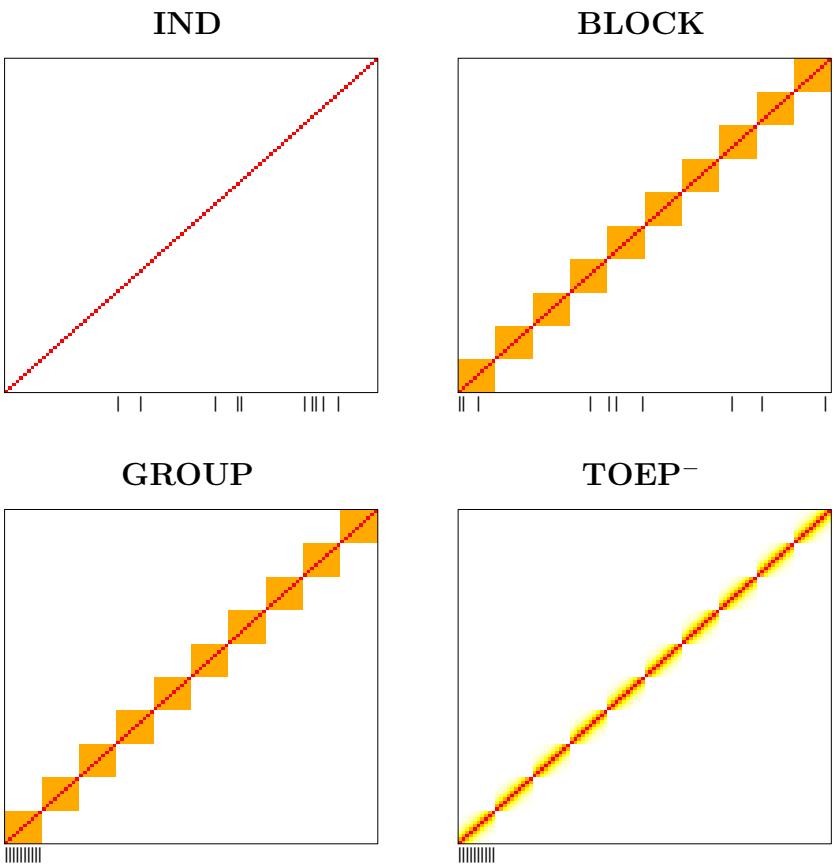


Figure 2.6 – Schematic representation of correlation structure for designs : IND (upper left), BLOCK (upper right), GROUP (bottom left) and TOEP^{-}(bottom right). Degrees of absolute correlation variyng from white - for weak correlation - to rhe red - for strong correlation. Ticks under the X-axis represent which variables have a non-null coefficient in $\boldsymbol{\beta}$.

2.6.2 Choice of Penalty Parameter

As it shown in the previous part, there exists many ways to define a consistent penalty for the ridge regression. These penalties (HOR, LOR) are commonly defined on OLS estimates. Thus, their computation is only possible when n is greater than p. Notice that a simple penalty $\frac{1}{n}$ is consistent too, even if it is not data specific. Commonly, the penalty is chosen by cross-validation to optimize the prediction error but it was not consistent. In the following, we always apply the

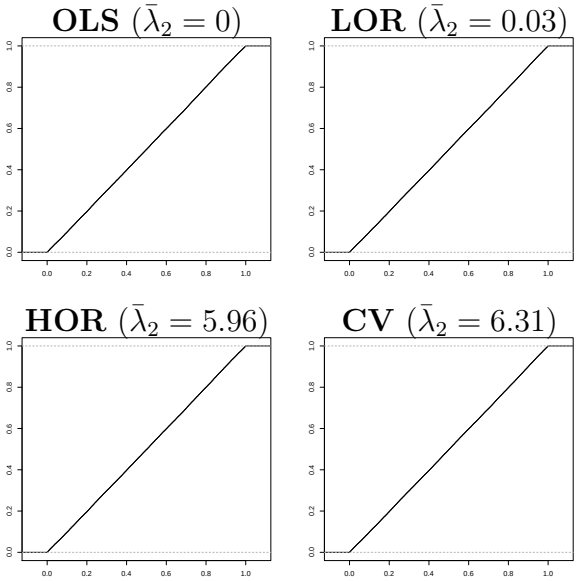


Figure 2.7 – Empirical Cumulative Distribution Function (ECDF) of estimated p-values for test under the null hypothesis with the B15 permutations test. Curves represent different level of penalty : 0 at the upper-left corresponding to the OLS reference ; LOR at the upper-right; HOR at the bottom-left and λ_2 chosen by CV at the bottom-right. Values between parenthesis represent the mean penalty for each definitions. Results are issued from 500 simulation on the IND design where $p = 100$, $n = 200$.

ridge regression with λ_2 varying between 10^{-5} and 10^3. This range allows to display some results of the various λ_2 estimations across all settings: when $n > p$, this ensures to have HOR, LOR and $\frac{1}{n}$ in the λ_2 grid; when $n < p$, this ensure us to display at least $\frac{1}{n}$ in the grid.

Figure 2.7 shows the empirical cumulative distribution function (ecdf) of p-values obtained on our permutation test for the Ols and the ridge regression with λ_2 estimated by HOR, LOR, or by Cross Validation. It exhibits no difference concerning the distribution of p-values while testing the null hypothesis. Our test seems to be valid whatever the penalty used if we compare it to the Ols reference. Moreover, penalty chosen by CV, when n is greater than p, is the HOR penalty in majority. This penalty seems to be consistent and optimal along all our simulations, when it is possible.

2.6.3 Results

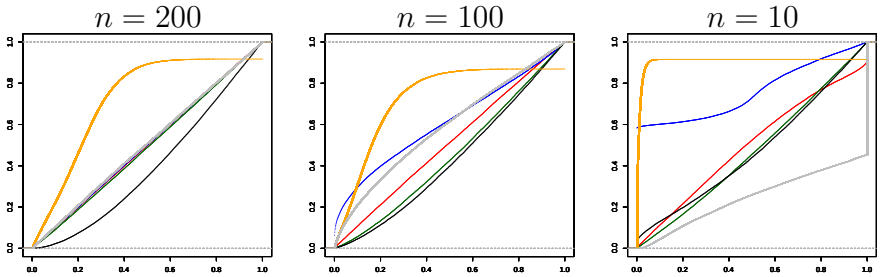


Figure 2.8 – Pvalue Empirical Cumulative Distribution Function (ECDF) of estimated p-values while testing the null hypothesis. Curves represent the following procedures : B15 in red; A99 in blue; M97 in green; C95 in black; classical F-test in orange and classical t-test in grey. Results are issued from 500 simulation on the IND design where $p = 100$, $n \in 10, 100, 200$ and λ_2 chosen by 10 folds CV.

P-value Distributions. Many test procedures have already been presented in the previous sections. Figure 2.8 shows the p-value ecdf of all procedures while testing null hypothesis for various sample size n when λ_2 is chosen by CV. For the most favorable case, when $n = 200$, all results follow the diagonal, except for the classical F-test, upper of the diagonal, and for the bootstrapping approach (C95), under of the diagonal. That induces, respectively, a good, a too weak and a too strong control of the type I error (FPR). For the other ratio n/p only procedures B15, M97 and C95 will have a good control of the FPR or a too stringent, at least, for M97 and C95.

As previously mentioned, the F-statistic estimated from ridge estimates does not follow a Fisher distribution and even if its distribution is close to a Fisher, there is no way to reliably estimate the corresponding parameters. This explain the ecdf upper the diagonal whatever the ratio n/p and confirms the aforementioned dangerousness to use it to test ridge estimates. Indeed, with an ecdf upper the diagonal the type I error will be higher than the expected level.

The ecdf for the classical T-test does surprisingly follow the diagonal for $n = 200$, but as we mentioned previously the CV chosen penalty often corresponds to the HOR estimation. Here, with a weak penalty and a large ratio n/p the bias is weak and the distribution of the T-

statistic will be close to a Student distribution. More precisely, it will be close to a normal distribution with regard to the degrees of freedom equal to 101 for the Ols in the same setting. Even if degrees of freedom are under-evaluated this has little effect on this special case, contrary to the Fisher distribution where an approximation on parameters can prove to be catastrophic. When n is lower than p, this is not longer favorable to the T-test and the performances are more erratic. In this last case, this approach seems highly inappropriate.

Robustness with Respect to Penalization. Figure 2.9 shows the evolution of type I and type II errors along a λ_2 grid where the penalty varying from 10^{-5} to 10^3. Similarly to previous results when $n > p$, all methods seem to control the FPR close to 5% with an exception for the K95 procedure, which shows FPR upper than the target level, and for the C95 procedure where the FPR is lower than α according to the ecdf results.

Methods based on residuals resampling (K95, F83 and A99) present along all results a FPR higher than expected. A possible explanation is that residuals are biased, especially for F83 and A99. Indeed for these methods, $\tilde{\mathbf{y}}$ is estimated from residuals stemming from biased coefficients. These coefficients are re-estimated with a new bias. The statistic, issued from permutations, is learned on values extremely biased when the original statistic suffer only from the original bias. These is a scale change between original statistic and statistics obtained by residuals permutations and this may explain the impossibility to control the FPR without removing the ridge estimate bias. To reduce the difference of scale, a tiny penalty could be used for permutations. In this case, penalties used for estimating the tested statistic and for estimating the distribution under the null hypothesis would not be the same.

When λ_2 increases then $\hat{\boldsymbol{\beta}}$ tends to 0 and residuals to $\mathbf{y}$. For F83, the resampling involves $\tilde{\mathbf{y}} = \hat{\mathbf{y}} + R$, so when λ_2 increases then $\hat{\mathbf{y}} \to 0$ and $R_{Y|X} \to \mathbf{y}$. In this case, this amounts to resampling $\mathbf{y}$ directly as

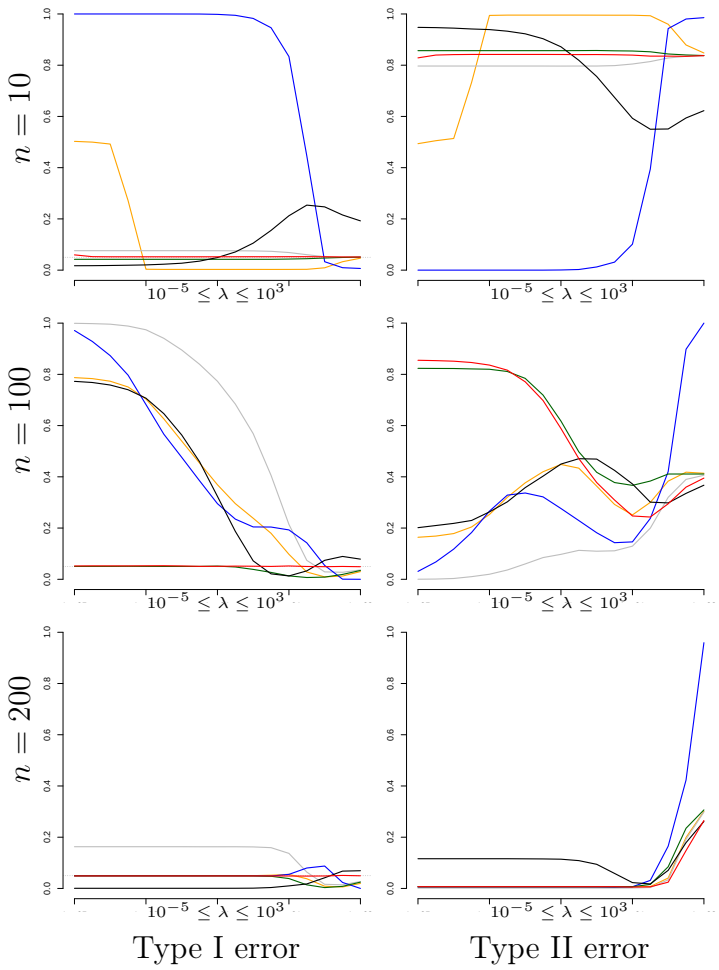


Figure 2.9 – Type I error (left column) and type II error (right column) for resampling procedures for varying penalties. Type I error is expected to be always equal to 5%, this threshold ie represented by the horizontal dotted grey line. For the type II error, the lower is the better. Curves represent following procedures : B15 in red; A99 in blue; M97 in green; C95 in black; F83 in orange and K95 in grey. Results are issued from 500 simulation on the IND design where $p = 100$, $n \in 10, 100, 200$ respectively at the upper, the middle and the bottom rows.

in M97 procedures. Figure 2.9 confirms this idea, where curve for M97 and F83 becomes similar for high value of λ_2. We can expect the same phenomenon with K95, because these three procedures have the same reference statistic r^2 and for a strong penalty all permuted measures converge to $\mathbf{y}$.

Only M97 and B15 methods seem to control type I error. These two methods are respectively based on a basic resampling of $\mathbf{y}$ and

of the tested variables. Here, the bias is present but is have the same scale of the bias of the reference statistic. As the scale is the same we can compare these statistics without inducing error and that explains, partly, the efficiency of these two methods.

Globally the C95 procedure, based on the bootstrap, over-controls the FPR with an interesting level of type II error. It is tempting to use this procedure, but the FPR, lower than α, it is not controlled. This can be explained by the high variance of $\hat{\beta}_j$ estimation for a null coefficient. The greater the variance, the closer to zero the decision statistic. In the other side, the bootstrap approach keeps a reasonable power because the variance of $\hat{\beta}_j$ for a non-null coefficient is less important. This differential comportment of the ridge for null or non-null coefficient is shown in the figure 2.10 where the variance to estimate null coefficient is very important.

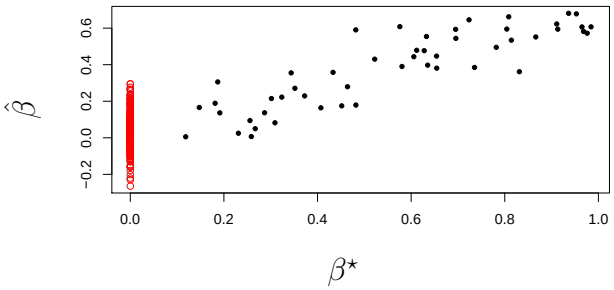


Figure 2.10 – Representation of ridge estimates $\hat{\boldsymbol{\beta}}$ versus real values $\beta^\star$ for a toy example with 250 variables whose 225 variables with null coefficients. Estimates of variables where $\beta^\star = 0$ are represented with red circle and by full black circle instead.

To conclude only our method (B15) and M97, which are similar in principle, allow an empirical control of the FPR at the expected α level. In our case the type II error is lower than Manly (2006) for an accurate FPR control.

Penalty Adjusted by Cross-Validation. Figure 2.11 shows the power and the FPR when λ_2 is chosen by CV, according to the usual protocol. These results confirm all the upper mentioned problems.

Table 2.5 – False Discovery Rate (FDR) for resampling procedures when λ_2 is chosen by CV. FDR is expected to be always equal to 5% with a BH procedure and it measured for following approaches : classical t-test; classical F-test; B15 ; A99; M97; C95; F83 and K95. Results are issued from 500 simulation on the IND, BLOCK, GROUP and TOEP$^-$designs where $p = 100$, $n \in 10, 100, 200$.

	IND			BLOCK			GROUP			TOEP$^-$		
	10	100	200	10	100	200	10	100	200	10	100	200
t-test	0.0	24.3	5.4	0.0	26.5	7.5	0.0	17.7	7.5	0.0	22.1	5.1
F-test	87.9	21.2	13.	84.6	12.5	11.	85.2	7.3	6.	87.1	22.1	15.6
B15	8.5	5.2	4.5	10.6	2.9	4.8	4.4	1.7	3.4	7.0	2.4	4.7
A99	14.2	46.7	6.	90.0	32.9	5.	90.0	3.0	0.	90.0	54.9	5.0
F83	50.2	14.5	5.6	79.4	8.7	5.3	50.8	6.4	6.3	76.0	20.9	5.3
K95	25.3	48.5	25.	49.5	42.6	33.	22.6	25.5	21.	34.9	62.9	33.3
M97	8.0	0.3	3.0	9.4	0.8	5.3	3.6	2.4	3.9	9.4	0.5	5.0
C95	80.8	0.6	0.	79.1	4.6	0.	7.6	14.9	6.	23.9	1.0	0.0

As expected, our testing procedure (B15) controls the FPR at 5% for all designs and ratio n/p. Our method is more impacted by the dimension of the problem compared to the correlation design, excepted for the BLOCK design where the maximal power is 80%. It is not surprising due to our sensibility to the masking of correlated relevant variables. However, our method seems to exhibit best or, at least, equal performance in all designs compared to the others when they show FPR control.

Table 2.5 shows the False Discovery Rate measures when it applied a BH procedure with $\alpha = 5\%$. Conclusions are the same compared to the analyze of the FPR but all methods, B15 and M97 included, seem not control the FDR when $n = 10$. This could be explained by the high variability of the FDP, especially when the problem is too difficult. The variability of the FDP and this impact on the FDR control will be discussed in the next chapter. Obviously, if a method does not control the FPR, then this method can not control the FDR.

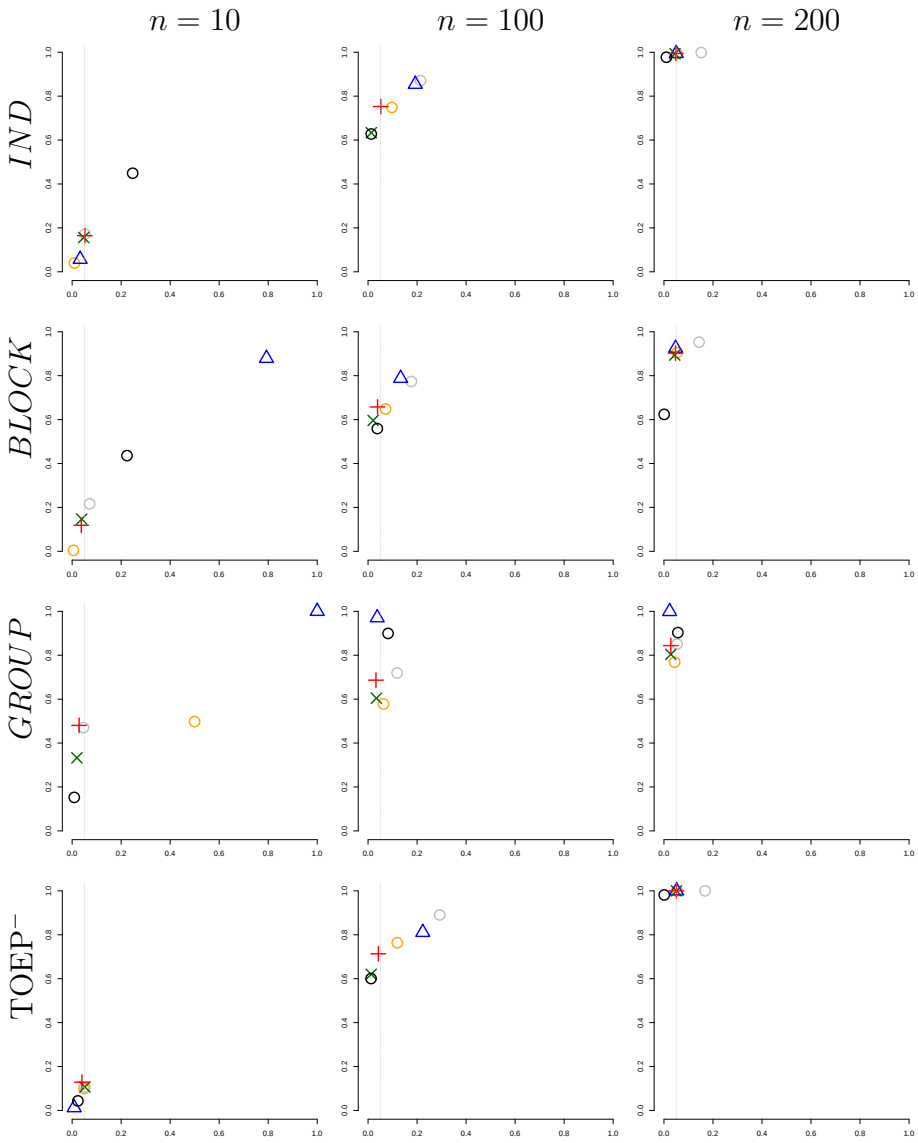


Figure 2.11 – Type I error (y-axis) and power (x-axis) for resampling procedures when λ_2 is chosen by CV. Type I error is expected to be always equal to 5%, this threshold is represented by the vertical dotted grey line. For the power, the upper is the better. Curves represent following procedures : B15 in red plus; A99 in blue triangle; M97 in green cross; C95 in black circle; F83 in orange circle and K95 in grey circle. Results are issued from 500 simulation on the IND, BLOCK, GROUP and $TOEP^-$designs where $p = 100$, $n \in 10, 100, 200$ respectively at the left, the middle and the right columns.

3. Two-Stage Approaches Based on Adaptive-Ridge Regression

This chapter presents the adaptive-ridge principle and its application to the "screen and clean" procedure (Wasserman and Roeder 2009). This procedure is based on data splitting (Cox 1975) with a portion to choose the hypothesis for test and the second portion for the evaluation of significance. This chapter is organized at following with a presentation of the original procedure of Wasserman and Roeder (2009), the adaptive-ridge principle with the improvement for the 'screen and clean" procedure and finally the transposition of this approach for estimation.

"Screen and clean" is a two-stage procedure proposed by Wasserman and Roeder (2009) to perform variable selection with statistical guarantees. The first stage screens variables to find a set of possibly relevant variables and the second stage cleans the set of candidate variables, thereby providing statistical guarantees on the risk of including irrelevant variables. The procedure considers a series of sparse models $\{\mathcal{F}_\lambda\}_{\lambda\in\Lambda}$, indexed by a parameter $\lambda \in \Lambda$, which may represent a penalty parameter for regularization methods or a size constraint for subset selection methods. The screening stage consists of two steps. In the first step, each model $\mathcal{F}_\lambda$ is fitted to (part of) the data, thereby selecting a set of possibly relevant variables, also known as the support of the model $\hat{\mathcal{S}}(\lambda) \subseteq \{1, \ldots, p\}$. Then, in the second step, a model se-

lection procedure chooses a single model $\mathcal{F}_{\hat{\lambda}}$ with its associated $\hat{\mathcal{S}}(\hat{\lambda})$. Next, the cleaning stage eliminates possibly irrelevant variables from $\hat{\mathcal{S}}(\hat{\lambda})$, resulting in the eventual set $\hat{\mathcal{S}}$ that provably controls the type I error rate. The original procedure relies on three independent subsamples of the original data $\mathcal{D}$, so as to ensure the consistency of the overall process. The following chart summarizes this procedure, showing the actual use of data $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$ that is made at each step:

$$\overbrace{\{1,\dots,p\} \xrightarrow[fit\ \ model]{step\ I(\mathcal{D}_1)} \{\hat{\mathcal{S}}(\lambda)\}_{\lambda\in\Lambda} \xrightarrow[select\ \ model]{step\ II(\mathcal{D}_1,\mathcal{D}_2)} \hat{\mathcal{S}}(\hat{\lambda})}^{screen} \underbrace{\xrightarrow[test\ \ support]{step\ III(\mathcal{D}_3)} \hat{\mathcal{S}}}_{clean}$$

Under suitable conditions, the screen and clean procedure performs consistent variable selection, that is, it asymptotically recovers the true support with probability one. The two main assumptions are that the screening stage should asymptotically avoid false negatives, and that the size of the true support should be constant, while the number of candidate variables is allowed to grow logarithmically in the number of examples. These assumptions are respectively described in rigorous terms by Meinshausen et al. (2009) as the "screening property" and "sparsity property".

Empirically, Wasserman and Roeder (2009) tested the procedure with the Lasso, univariate testing, and forward stepwise regression at step I of the screening stage. At step II, model selection was always based on ordinary least squares (OLS) regression. The OLS parameters were adjusted on the "training" subsample $\mathcal{D}_1$, using the variables in $\{\hat{\mathcal{S}}(\lambda)\}_{\lambda\in\Lambda}$, and model selection consisted in minimizing the empirical error on the "validation" subsample $\mathcal{D}_2$ with respect to λ. Cleaning was then finally performed by testing the nullity of the OLS coefficients using the independent "test" subsample $\mathcal{D}_3$. Like is shown in Fithian et al. (2015) the type I error of this procedure is not

$$\mathbb{P}(\text{reject } \mathcal{H}_0 | \mathcal{H}_0 \text{true})$$

but is

$$\mathbb{P}(\text{reject } \mathcal{H}_0 | \mathcal{H}_0 \text{true}, \hat{\mathcal{S}}(\hat{\lambda})).$$

That must be taken into account to the results interpretation, where an causal element to be selected must be observed and pre-selected in $\hat{\mathcal{S}}(\hat{\lambda})$. Wasserman and Roeder (2009) conclude that the variants using multivariate regression (Lasso and forward stepwise) have similar performances, way above univariate testing.

3.1 Adaptive-Ridge for Inference

We now introduce the improvements that we propose at each stage of the process. Our main methodological contribution lies at the cleaning stage, but we introduce other modifications at the screening stage that have considerable practical outcomes.

3.1.1 Cleaning Stage

The original cleaning stage of Wasserman and Roeder (2009) is based on the ordinary least square (OLS) estimate. This choice is amenable to efficient exact testing procedure for selecting the relevant variables, where the false discovery rate can be provably controlled. However, this advantage comes at a high price:

— first, the procedure can only be used if the OLS is applicable, which requires that the number of variables $\left|\hat{\mathcal{S}}(\hat{\lambda})\right|$ that passed the screening stage is smaller than the number of examples $|\mathcal{D}_3|$ reserved for the cleaning stage;

— second, the only information retained from the screening stage is the support $\hat{\mathcal{S}}(\hat{\lambda})$ itself. There are no other statistics about the estimated regression coefficients that are transferred to this stage.

We propose to make a more effective use of the data reserved for the screening stage, by retaining the magnitude of the regression coefficients $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ obtained at this stage. Our procedure allows for a cleaning stage in the high-dimensional setup (that is, $\left|\hat{\mathcal{S}}(\hat{\lambda})\right| \gg |\mathcal{D}_3|$), and our experiments show that conveying the magnitude of the regression co-

efficients $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ to the cleaning stage systematically improves the power of the procedure: the statistics produced result in dramatic increases in sensitivity (that is, in true positives) at any false discovery rate (see Figure 3.3).

Technically, the magnitude of the regression coefficients $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ is forwarded to the cleaning stage via an adaptive-ridge penalty term. Adaptive refers here to the adaptation of the penalty terms to the data at hand. The penalty shape is adjusted to the "training" subsample $\mathcal{D}_1$, its strength is set thanks to the "validation" subsample $\mathcal{D}_2$, and it is finally applied to cleaning stage on $\mathcal{D}_3$. This process is detailed below.

Computing the Regression Coefficients Our cleaning stage is specifically designed for a screening stage based on the Lasso or more generally on the elastic-net estimator (Zou and Hastie 2005), which is nowadays widely used to tackle simultaneously variable estimation and selection. In our framework, it also offers the possibility to select larger supports at the screening stage, thus favoring the "screening property" (more details will be given in Section 3.1.2). We recall that selecting a large support is problematic when the cleaning stage relies on the OLS, which may then be unstable, or even ill-defined. We avoid this problem by using penalization at both stages of the feature selection method.

Our approach relies on an alternative view of the elastic-net, seen as an adaptive-ℓ_2 penalization scheme (Grandvalet 1998, Grandvalet and Canu 1999). This view is formalized by a variational form of the elastic-net:

$$\begin{aligned} \min_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\tau}\in\mathbb{R}^p} \; & \{J(\boldsymbol{\beta}) + \sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda_1}{\tau_j} + \lambda_2\right)\} \\ \text{s. t. } & \sum_{j=1}^{p} \tau_j - \sum_{j=1}^{p} |\beta_j| \leq 0 \;\; , \;\; \tau_j \geq 0 \;\; , \;\; j = 1, \ldots, p \;\; . \end{aligned} \tag{3.1}$$

The variable $\boldsymbol{\tau}$ introduced in this formulation, which adapts the ℓ_2 penalty to the data, can be shown to lead to the following adaptive-

ridge penalty:

$$\sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda_1}{|\hat{\beta}_j(\lambda)|} + \lambda_2 \right) \quad , \tag{3.2}$$

where the coefficients $\hat{\beta}_j(\lambda)$ are the solution to the elastic-net problem (1.2.4).

Using this adaptive-ℓ_2 penalty returns the original elastic-net estimator, as shown in the following lemma.

Lemma 1. *The quadratic penalty in $\boldsymbol{\beta}$ in* (3.1) *acts as the elastic-net penalty $\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$.*

Proof. The Lagrangian of Problem (3.1) is:

$$L(\boldsymbol{\beta}) = J(\boldsymbol{\beta}) + \sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda_1}{\tau_j} + \lambda_2 \right) + \nu_0 \left(\sum_{j=1}^{p} \tau_j - \|\boldsymbol{\beta}\|_1 \right) - \sum_{j=1}^{p} \nu_j \tau_j \quad .$$

Thus, the first order optimality conditions for τ_j are

$$\begin{aligned} \frac{\partial L}{\partial \tau_j}(\tau_j^\star) = 0 \Leftrightarrow -\frac{\lambda_1 \beta_j^2}{\tau_j^{\star 2}} + \nu_0 - \nu_j = 0 \\ \Leftrightarrow -\lambda_1 \beta_j^2 + \nu_0 \, \tau_j^{\star 2} - \nu_j \, \tau_j^{\star 2} = 0 \\ \Rightarrow -\lambda_1 \beta_j^2 + \nu_0 \, \tau_j^{\star 2} = 0 \quad , \end{aligned}$$

where the term in ν_j vanishes due to complementary slackness, which implies here $\nu_j \tau_j^\star = 0$. Together with the constraints of Problem (3.1), the last equation implies $\tau_j^\star = |\beta_j|$, hence Problem (3.1) is equivalent to

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} J(\boldsymbol{\beta}) + \left(\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 \right) \quad ,$$

which is the original elastic-net formulation. □

This equivalence will be used here for defining a data-dependent penalty, determined in the screening stage, that will also be applied in the cleaning stage. In this process, our primary aim is to retain the magnitude of the coefficients of $\hat{\boldsymbol{\beta}}(\hat{\lambda})$ in addition to the support $\hat{\mathcal{S}}(\hat{\lambda})$: the small coefficients of the screening stage will be encouraged to be small in the cleaning stage, whereas the largest ones will be less penalized. The expected side-effects of this penalization at the cleaning

stage are to allow for the processing of more variables and to stabilize the estimation procedure.

Cleaning is eventually performed by testing the nullity of the adaptive-ridge coefficients using the independent "test" subsample $\mathcal{D}_3$. The statistics computed from our penalized cleaning stage improve the power of the procedure: we observe a dramatic increase in sensitivity (that is, in true positives) at any false discovery rate (see Figure 3.3 of the numerical experiment section). However, using penalized estimators raises a difficulty for the calibration of the statistical tests derived from these statistics. We propose to resolve this issue through the use of the permutation tests presented in Section 2.4.2.

Testing the Coefficients The main goal of the cleaning stage is to select variables, for these which have passed the screening (that is, in $\hat{\mathcal{S}}(\hat{\lambda})$), with a control of the FDR. To test variables we propose to use a permutation based F-test but the knockoff approach can be used instead.

F-test For each $j \in \hat{\mathcal{S}}(\hat{\lambda})$ we apply our permutation test based on the F-statistic (2.17). In Chapter 2, we dealt with hypothesis testing for the standard ridge regression problem where the penalty λ is identical for all variables. For adaptive-ridge, the penalization matrix $\lambda\mathbf{I}$ is replaced by a diagonal matrix whose diagonal elements are given in the right-hand-side of (3.2). The final support is defined by the set F indexing for which the null hypothesis is rejected, that is, $\hat{\mathcal{S}} = \{j \in \hat{\mathcal{S}}(\hat{\lambda}) : P_j \leq \alpha\}$, where the p-values P_j are corrected by the Bonferroni or the Benjamini-Hochberg procedures to control the FWER or the FDR respectively.

FDR Control The Benjamini-Hochberg procedure ensures a FDR $\leq \alpha$. More precisely the FDR is equal to $\alpha\pi_0$, if all p-values are independent, as it shown in equation (1.10). The proportion of true null hypothesis, denoted π_0, is unknown, but it is commonly consid-

ered to be close to 1 assuming that true null hypothesis outnumber true alternative hypotheses. In the screen and clean procedure π_0 is the proportion of truly irrelevant variables in $\hat{\mathcal{S}}(\hat{\lambda})$, that is the false discovery proportion following from the screening stage. This quantity is unknown, except in simulations, but the assumption of $\pi_0 \simeq 1$ does not hold. Hence, the Benjamini-Hochberg procedure induces an overly stringent control of FDR. The LBE method (Dalmasso et al. 2005) estimates π_0 on $\hat{\mathcal{S}}(\hat{\lambda})$, so as to control the FDR level closer to the prescribed α.

Note that Wasserman and Roeder (2009) not used correction procedure on p-values for multiple testing, arguing that the method is conservative provided the screening is severe enough following his results. This is particularly true in their setup where they arbitrarily limit the size of $\hat{\mathcal{S}}(\hat{\lambda})$ at $\sqrt{n}$.

Knockoff-Test The screening step used with the Lasso ensure that $n_{\mathcal{D}_2} > |\hat{\mathcal{S}}(\hat{\lambda})|$. The knockoff procedure (Barber and Candes 2014), can be applied on $\mathbf{X}_{\mathcal{D}_2}$ limited by $|\hat{\mathcal{S}}(\hat{\lambda})|$ to create the augmented matrix $\left[\mathbf{X}_{\mathcal{D}_2}\tilde{\mathbf{X}}_{\mathcal{D}_2}\right]$. Theoretically , we can test each variables, of this augmented matrix, with specifics statistics measured on estimates of Lasso, ridge (adaptive or not) or OLS .

3.1.2 Screening Stage

Wasserman and Roeder (2009) propose to use two subsamples at the cleaning stage in order to establish the consistency of the screen and clean procedure. Indeed, this consistency relies partly on the fact that all relevant variables pass the screening stage with very high probability. This "screening property" Meinshausen et al. (2009) was established using the protocol described in Section 3. To our knowledge, it remains to be proved for model selection based on cross-validation. However, Wasserman and Roeder (2009) mention another procedure relying on two independent subsamples of the original data $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, where model selection relies on leave-one-out cross-validation on $\mathcal{D}_1$ and $\mathcal{D}_2$

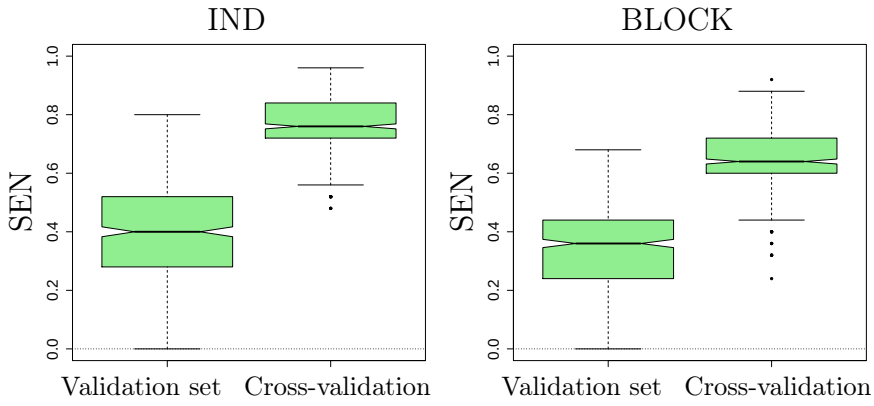


Figure 3.1 – Sensitivity of the screen and clean procedure (the higher, the better), for the two model selection strategies at the screening stage, and FDR controlled at 5% based on the permutation test. Lasso regression is used in the screening stage and adaptive-ridge regression in the cleaning stage. Each boxplot is computed based over 500 replications for the IND and BLOCK simulation designs, with $n = 250$, $p = 500$, $|\hat{\mathcal{S}}^*| = 25$ and $\rho = 0.5$ (see Section 2.6.1 for full description).

is reserved for cleaning. The following chart summarizes this modified procedure:

$$\overbrace{\{1,\dots,p\} \xrightarrow[fit\ \ model]{step\ I(\mathcal{D}_1)} \{\hat{\mathcal{S}}(\lambda)\}_{\lambda\in\Lambda} \xrightarrow[select\ \ model]{step\ II(\mathcal{D}_1)} \hat{\mathcal{S}}(\hat{\lambda})}^{screen} \underbrace{\xrightarrow[test\ \ support]{step\ III(\mathcal{D}_2)} \hat{\mathcal{S}}}_{clean}$$

Hence, half of the data are now devoted to each stage of the method. We followed here this variant, which results in important sensitivity gains for the overall selection procedure, as illustrated in Figure 3.1.

We slightly depart from Wasserman and Roeder (2009), by selecting the model by 10-fold cross-validation, and, more importantly, by using the sum of squares residuals of the *penalized* estimator for model selection. Note that Wasserman and Roeder (2009), and later Meinshausen et al. (2009) based model selection on the OLS estimate using the support $\hat{\mathcal{S}}_\lambda$. This choice results in an implicit limitation of the size of the selected support $|\hat{\mathcal{S}}(\hat{\lambda})| < \frac{n}{2}$, which is actually implemented more severely as $|\hat{\mathcal{S}}(\hat{\lambda})| \leq \sqrt{n}$ and $|\hat{\mathcal{S}}(\hat{\lambda})| \leq \frac{n}{6}$ by Wasserman and Roeder (2009) and Meinshausen et al. (2009) respectively. Our model selection criterion allows for more variables to be transferred at the

cleaning stage, so that the screening property is more likely to hold. Our complete protocol, based on the elastic-net at the screening stage can be summarized as follows (the Lasso case corresponds to setting $\lambda_2 = 0$):

Input: dataset $\mathcal{D}$ comprising the $n \times p$ covariate matrix $\mathbf{X}$ and the n-dimensional response vector $\mathbf{y}$, Λ_1 a set of trial values for the $l1$ penalty, Λ_2 a set of trial values for the $l2$ penalty, and B the number of permutations

Output: index $\hat{\mathcal{S}}$ of variables selected during screening, and P a vector with p-values associated to these variables

0. Split randomly $\mathcal{D}$ in two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$ of size $n/2$
1. "Screening" on $\mathcal{D}_1$:
   (a) Select $(\hat{\lambda}_1, \hat{\lambda}_2)$ in $\Lambda_1 \times \Lambda_2$ by 10-fold cross-validation
   (b) Define $\hat{\mathcal{S}} = \{j \in \{1, \ldots, p\} : \hat{\beta}_j(\hat{\lambda}_1, \hat{\lambda}_2) \neq 0\}$
2. "Cleaning" on $\mathcal{D}_2$:
   (a) Compute AR estimate $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}(\hat{\lambda}_1, \hat{\lambda}_2)$ from $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}(\hat{\lambda}_1, \hat{\lambda}_2)$
   (b) Compute RSS^{Ω}, the residual sum of squares for $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}$ on $\mathcal{D}_2$
   (c) for $j \in \hat{\mathcal{S}}$:
   - Compute AR estimate $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_{\setminus j}}(\hat{\lambda}_1, \hat{\lambda}_2)$
   - Compute RSS^{ω} for $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}_{\setminus j}}$ on $\mathcal{D}_2$
   - $F = \frac{RSS^{\omega} - RSS^{\Omega}}{RSS^{\Omega}}$
   - for $b \in \{1, \ldots, B\}$:
     — $\mathbf{X}^{(b)}_{\mathcal{D}_2 \cdot} = \mathbf{X}_{\mathcal{D}_2 \cdot}$.
     — Permute randomly the column j of $\mathbf{X}^{(b)}_{\mathcal{D}_2 \cdot}$.
     — Compute AR estimate $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}{}^{(b)}(\hat{\lambda}_1, \hat{\lambda}_2)$ on $\mathbf{X}^{(b)}_{\mathcal{D}_2 \cdot}$.
     — Compute $RSS^{\Omega(b)}$ for $\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}}{}^{(b)}$
     — $F^{(b)} = \frac{RSS^{\omega} - RSS^{\Omega(b)}}{RSS^{\Omega(b)}}$
   - Compute the empirical p-value $P_j = \frac{1}{B}\#\{F \leq F^{(b)}\}$
   (d) Multiple testing adjustment on p-values

3.1.3 Analysis of the Orthonormal Design

Here, we propose a detailed analysis of the benefits of the procedure in the orthonormal design, where we assume that we have 2 samples of size n with design matrices $n^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X} = \mathbf{I}$. In this situation, the

screening stage based on the Lasso provides

$$\hat{\beta}_j^{\text{screen}}(\lambda) = \left(1 - \frac{\lambda}{n|\hat{\beta}_j^{\text{ols}}|}\right)_+ \hat{\beta}_j^{\text{ols}} \quad ,$$

where $\hat{\boldsymbol{\beta}}(0)$ is the ordinary least squares estimator (Tibshirani 1996). Assuming additionally a Gaussian noise in the model, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, the probability that variable j does not pass the screening stage is:

$$\begin{aligned} P(\beta_j^\star, \lambda) &= \mathbb{P}\left[\hat{\beta}_j^{\text{screen}}(\lambda) = 0\right] \\ &= \Phi\left(n^{1/2}\sigma^{-1}(n^{-1}\lambda - \beta_j^\star)\right) - \Phi\left(-n^{1/2}\sigma^{-1}(n^{-1}\lambda + \beta_j^\star)\right) \quad , \end{aligned}$$

where Φ is the cumulative distribution of the standard normal distribution. Then, as the cleaning stage operates on an independent sample, the distributions for the cleaning stage estimators are:

$$f(\hat{\beta}_j^{\text{clean}}) = P(\beta_j^\star, \lambda)\, \delta(\hat{\beta}_j^{\text{clean}}) + (1 - P(\beta_j^\star, \lambda))\, g(\hat{\beta}_j^{\text{clean}}) \quad ,$$

with

$$g(\hat{\beta}_j^{\text{clean}}) = n^{1/2}\sigma^{-1}\varphi\left(\frac{n^{1/2}}{\sigma}(\hat{\beta}_j^{\text{clean}} - \beta_j^\star)\right)$$

for the OLS estimator, and

$$g(\hat{\beta}_j^{\text{clean}}) = n^{1/2}\sigma^{-1}\int_0^1 x^{-1}\varphi\left(x^{-1}n^{1/2}\sigma^{-1}(\hat{\beta}_j^{\text{clean}} - \beta_j^\star)\right)\, h_j(x)\, dx$$

for the adaptive-ridge (AR) estimator, where δ is the Dirac delta function, φ is the standard normal distribution, and h_j is the distribution of the shrinkage coefficient that is applied to variable j by AR, that is, the distribution of $|\hat{\beta}_j^{\text{screen}}|/(|\hat{\beta}_j^{\text{screen}}| + n^{-1}\lambda)$. There is no simple analytical form of the overall distribution for the adaptive-ridge estimator, but these formulas are however interesting, in that they exhibit the three parameters of importance for the distribution of the cleaning estimators, namely $\beta_j^\star$, $n^{1/2}\sigma^{-1}$, and $n^{-1}\lambda$. Furthermore, in all expressions, $n^{1/2}\sigma^{-1}$ acts as a scale parameter. We can therefore provide a scale-free analysis by studying the role of the normalized penalty parameter $n^{-1/2}\sigma^{-1}\lambda$ on the normalized estimator $n^{1/2}\sigma^{-1}\hat{\beta}_j^{\text{clean}}$, when the normalized true parameter $n^{1/2}\sigma^{-1}\beta_j^\star$ varies.

We now make use of these observations to compare the power of the statistical testing of the nullity of $\beta_j^\star$ from the OLS and from the AR cleaning estimator. First, the expected type-I-error is fixed to 1% using the distributions of $\hat{\beta}_j^{\text{clean}}$ for the OLS and the AR estimator for $\beta_j^\star = 0$, and we then compute the expected type-II-error according to $n^{1/2}\sigma^{-1}\beta_j^\star$. A significance level of 1% roughly corresponds to the effective significance level for unitary tests in our experiments of Section 3.1.4 aiming at controlling the FDR at 5% using the Benjamini-Hochberg procedure.

Figure 3.2 represents graphs spanning the possible values of $n^{-1/2}\sigma^{-1}\lambda$. For readability, we indexed subfigures by $P(0,\lambda)$, the probability that

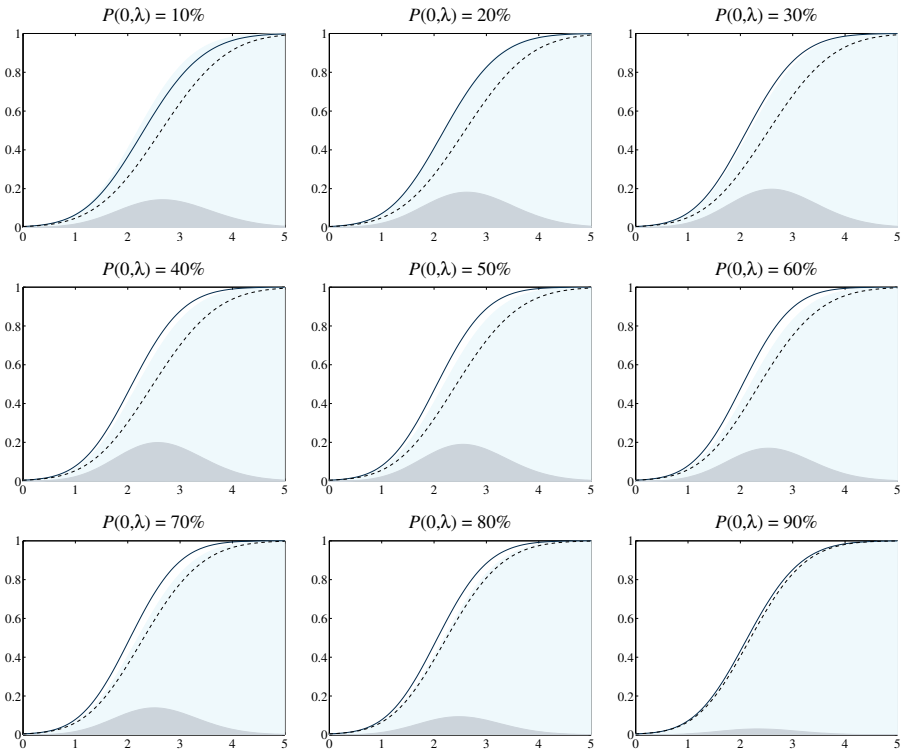


Figure 3.2 – Power (or sensitivity) as a function of $n^{1/2}\sigma^{-1}\beta_j^\star$, for a univariate test at the 1% level based on OLS cleaning (dashed) or AR cleaning (plain). The light gray area in the bottom displays the difference between the two curves, and the boundary of the very light blue area, included for cross-reference, represents the best result achieved using the OLS cleaning estimator, for $P(0,\lambda) = 90\%$.

a null variable is filtered at screening stage. We observe that, for any λ value, the test based on the AR estimator has uniformly higher power

than the one based on the OLS estimator. Furthermore, for most λ values, AR cleaning performs better than the best λ setting for OLS cleaning. This means that AR cleaning often brings more than having an oracle for selecting λ at the screening stage.

3.1.4 Results

In the following, we discuss the IND, BLOCK, GROUP and TOEP$^-$ designs (presented in chapter 2) with $n \in \{250, 500\}$, $p \in \{250, 500\}$, $|\mathcal{S}^\star| = 25$, $\rho = 0.5$, SNR $\in \{0.5, 4, 8\}$ and a size block equal to 25, since the relative behavior of all methods is representative of the general pattern that we observed across all simulation settings. These setups lead to feasible but challenging problems for model selection. All variants of the screen and clean procedure are evaluated here with respect to their sensitivity (SEN) (1.3), for a controlled false discovery rate FDR (1.7).

We conducted experiments with the elastic-net and the Lasso at the screening stage. Although they may result in different set sizes at the output of the screening stage (see Table 3.1), they eventually end up to identical results after the cleaning stage, in terms of FDR control and sensitivity. In what follows, we thus only report results with the simplest Lasso screening stage that relies on a single penalty parameter.

Table 3.1 – Average number of variables passing the Lasso and elastic-net screening stages, computed over 500 simulations for each design, for $n = 250$ and $p = 500$ where estimates are done on $\frac{n}{2}$.

Simulation design	IND	BLOCK	GROUP	TOEP$^-$
Lasso	96.95	89.51	39.77	75.80
Elastic-net	97.11	91.72	58.40	86.26

We first show the importance of the cleaning stage for FDR control. We then show the benefits of our proposal compared to the original

Table 3.2 – False discovery rate FDR and sensitivity SEN, computed over 500 simulations for each design where $n = 250$ and $p = 500$. The screening stage is not calibrated; the cleaning stage, performed with adaptive ridge, is calibrated to control the FDR below 5%, using the Benjamini-Hochberg procedure.

SNR	Step	IND		BLOCK		GROUP		TOEP$^-$	
		FDR	SEN	FDR	SEN	FDR	SEN	FDR	SEN
0.5	Screening	50.8	9.8	67.3	11.4	56.3	23.8	49.5	6.7
	Cleaning	4.6	0.4	10.6	0.5	2.8	4.3	2.5	0.5
4	Screening	76.7	87.5	76.0	83.9	38.9	86.2	79.9	56.5
	Cleaning	4.2	76.1	2.8	64.8	1.7	37.7	4.3	39.6

procedure of Wasserman and Roeder (2009) and to the univariate approach. At last, we discuss about the knockoff test (Barber and Candes 2014) used instead our permutation test, the FDR variability, the importance of true null hypothesis ration in $\hat{\mathcal{S}}(\hat{\lambda})$ and the ranking difference between estimates and p-values. The variable selection method of Lockhart et al. (2014) was not included in these experiments, because it did not produce convincing results in these small n large p designs where the noise variance is not assumed to be known.

Importance of the Cleaning Stage Table 3.2 shows that the cleaning step is essential to control the FDR at the desired level. In the screening stage, the variables selected by the Lasso are way too numerous: first, the penalty parameter is determined to optimize the cross-validated mean squared error, which is not optimal for model selection; second, we are far from the asymptotic regime where support recovery can be achieved. As a result, the Lasso performs rather poorly.

For most cases, cleaning enables the control of the FDR, leading of course to a decrease in sensitivity, which is moderate for independent variables, and higher in the presence of correlations. The BLOCK design with $SNR = 0.5$ provides a noticeable exception to this general rule, with FDR twice higher than prescribed. Covariate masking is especially important for this design, where relevant and irrelevant vari-

ables are correlated. In this situation, truly irrelevant variables may become relevant at the screening stage: when a relevant covariate is not included in $\hat{\mathcal{S}}(\hat{\lambda})$. In this case, the response variable appears to depend directly of all the irrelevant covariates that are correlated to this missing covariate. Following the disappearance of the truly relevant covariate, these irrelevant covariates enter on the Markov blanket of the response variable (see Chapter 1). If we define true positives relative to this Markov blanket at screening stage, the FDR is equal to 1% instead to 10.6%.

Comparisons of Controlled Selection Procedures Figure 3.3 provides a global picture of sensitivity according to FDR, for the test statistics computed in the cleaning stage. First, we observe that the direct univariate approach, which simply considers a t-statistic for each variable independently, is by far the worst option in the IND, BLOCK and TOEP$^-$ designs, and by far the best in the GROUP design. In this last situation, the univariate approach confidently detects all the correlated variables of the relevant group, while the regression-based approaches are hindered by the high level of correlation between variables. The TOEP$^-$ design present the same structure with all relevant variables in the same block, but with different level of correlation compared to the GROUP design. However, univariate work less than other for TOEP$^-$ design due to negative correlation between covariates. Betting on the univariate approach may thus be profitable, but it is also risky due to its extremely erratic behavior. Second, we see that our adaptive-ridge cleaning systematically dominates the original OLS cleaning. To isolate the effect of transferring the magnitude of weights from the effect of the regularization brought by adaptive-ridge, we show the results obtained from a cleaning step based on plain ridge regression (with regularization parameter set by cross-validation). We see that ridge regression cleaning improves upon OLS cleaning, but that adaptive-ridge cleaning brings this improvement much further, thus confirming the value of the weight transfer from the screening stage to

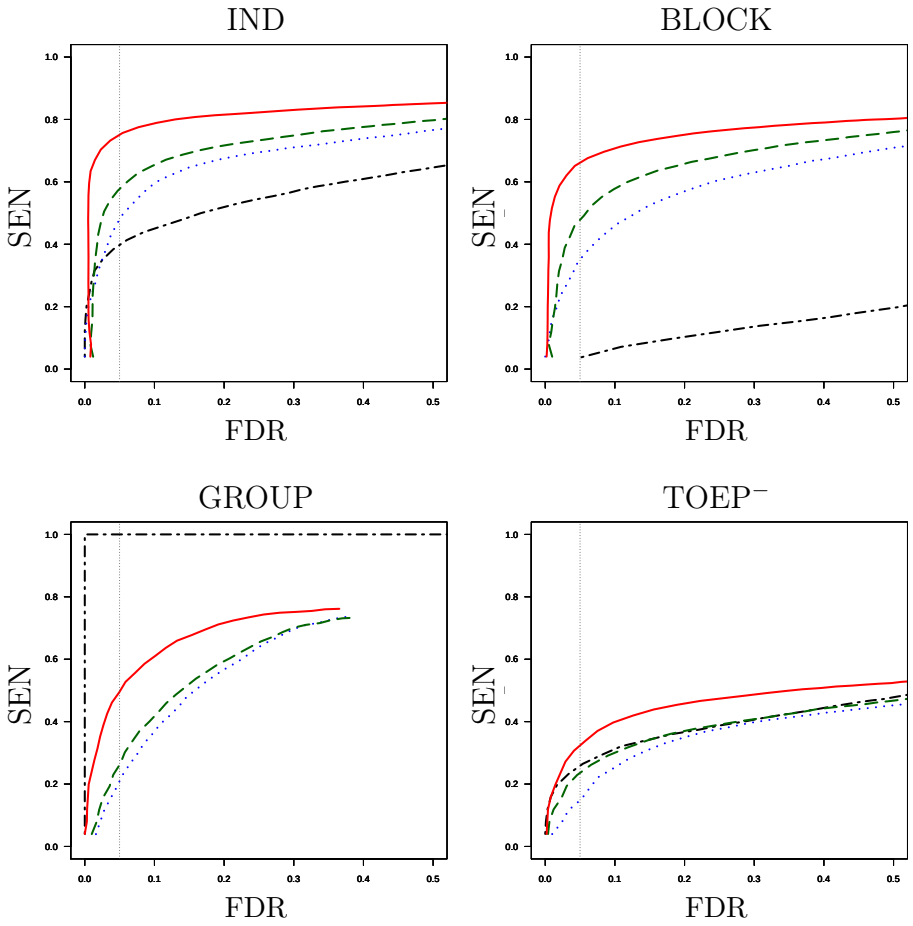


Figure 3.3 – Sensitivity SEN versus False Discovery Rate FDR (the higher, the better). Lasso screening followed by: adaptive-ridge cleaning (red solid line), ridge cleaning (green dashed line), OLS cleaning (blue dotted line), univariate testing (black dot-dashed line). All curves are indexed by the rank of the test statistics, and averaged over the 500 simulations of each design for $n = 250$, $p = 500$ and SNR $= 4$. The vertical dotted line marks the 5% FDR level.

the cleaning stage.

Better results for ridge or adaptive-ridge on correlated dataset are expected because the ridge regression performs better than the OLS in presence of correlation. However, results for the IND design appears counter-intuitive, because the variables in $\hat{\mathcal{S}}$ are not correlated. That could be explain by a lack of power for the t-test based on OLS estimates. Indeed, degrees of liberty of t-test is equal to $n - p$ and it $\hat{\mathcal{S}}(\hat{\lambda})$ is close to $n - p$ then the power of this test will be impacted.

Using the arbitrary threshold on the size of $\hat{\mathcal{S}}(\hat{\lambda})$ at $\sqrt{n}$ (Wasserman and Roeder 2009) is a manner to avoid this loss of degrees of liberty. In this case, the degrees of freedom $n - p$, for the student-test in the OLS, grew three times compared to the degrees of freedom for the unthreshold $\hat{\mathcal{S}}(\hat{\lambda})$. To conclude, the ridge regression support better the high-dimensional case and our permutation test is not constrained to the degrees of freedom contrary to the exact test for the OLS. That explains the upper curve for the ridge regression than the OLS.

Table 3.3 shows the actual operating conditions of the variable selection procedures, when a threshold on the test statistics has to be set to control the FDR. Here, the threshold is set to control the FDR at a 5% level, using Benjamini-Hochberg correction. Due to the variability of FDP (see Table 3.4), we consider a good control of FDR if the expected level α (5%) is contained on the confidence interval of the FDP at 95%.

This control (see Table 3.3) is globally effective for the screen and clean procedures, except for variable selection based on univariate testing in the BLOCK design. Univariate approach (see Chapter 1) measure the marginal correlation between each explicative variable and the response. Previously, we have proposed specific definition of FDR when a relevant variable is missing on a block. In that sense for the univariate approach, all irrelevant variables became relevant and the FDR became lower than 5%. We had the same conclusion for the BLOCK design when $SNR = 0.5$ when all methods appear to not control the FDR based on the basic definition of FP. That could be due to the difficulty to distinguish the specificity of variables really under $\mathcal{H}_1$ to their co-variables under $\mathcal{H}_0$ when the noise is too high.

In all designs, our proposal dramatically improves over the original OLS strategy, with sensitivity gains ranging from 50% to 100%. All differences in sensitivity are statistically significant. This is illustrated by the boxplots of Figure 3.4, which represent the difference of sensitivity between our Lasso-based screen and clean procedure and itss competitors. Notches bellow zero indicate that the competitor has

Table 3.3 – False discovery rate FDR and sensitivity SEN, computed over 500 simulations for each design. The cleaning stage is calibrated to control the FDR below 5%, using the Benjamini-Hochberg procedure. Our adaptive-ridge (AR) cleaning is compared with the original (OLS) cleaning, the intermediate ridge cleaning (RIDGE), and univariate testing (UNIVAR). The best sensitivity for an effective control of the FDR (5% within the 95% confidence interval) is shown in bold. Failures of FDR control are shown in red.

	n	p	SNR	AR		RIDGE		OLS		UNIVAR	
				FDR	SEN	FDR	SEN	FDR	SEN	FDR	SEN
IND	250	250	0.5	2.8	0.7	4.0	**0.8**	2.4	0.5	3.3	2.6
			4	3.9	**82.5**	4.1	72.0	3.8	68.1	4.6	45.8
			8	3.9	**96.7**	3.4	89.6	3.4	91.7	4.6	47.6
		500	0.5	4.6	**0.4**	4.7	0.3	2.9	0.2	3.9	1.3
			4	4.2	**76.1**	4.6	57.9	3.9	48.3	4.4	40.4
			8	3.9	**93.3**	3.7	75.8	3.2	80.7	4.4	42.1
	500	250	0.5	5.1	**4.1**	5.6	**4.1**	3.8	3.3	4.1	12.3
			4	3.5	**94.7**	3.6	93.5	3.6	93.3	4.5	66.4
			8	3.9	**99.8**	3.7	99.6	3.6	99.6	4.6	67.2
		500	0.5	4.1	**3.1**	3.9	3.0	3.3	2.4	4.6	8.9
			4	3.7	**94.0**	3.9	91.2	3.7	91.6	4.7	63.1
			8	3.9	**99.7**	4.0	98.5	3.6	99.1	4.6	64.1
BLOCK	250	250	0.5	12.9	0.7	7.9	0.5	6.4	**0.4**	67.4	25.2
			4	2.5	**64.8**	3.5	57.4	4.0	49.2	84.5	85.6
			8	1.8	**88.8**	3.2	83.0	3.3	82.3	84.7	86.5
		500	0.5	10.6	0.5	8.5	0.4	6.9	**0.3**	53.2	9.4
			4	2.8	**64.8**	3.6	49.8	3.1	37.1	86.4	71.0
			8	2.3	**88.7**	4.0	76.7	3.6	75.8	86.6	75.2
	500	250	0.5	14.5	2.5	8.6	1.4	7.8	1.2	79.6	59.3
			4	1.7	**83.8**	3.1	81.7	3.1	80.7	86.5	93.6
			8	1.5	**96.8**	3.3	95.6	3.1	95.6	86.8	93.9
		500	0.5	12.4	2.6	9.1	1.7	7.4	1.4	77.3	86.2
			4	2.3	**86.6**	3.8	82.8	3.8	81.9	89.5	86.5
			8	1.9	**97.9**	3.7	95.9	3.9	96.2	89.7	87.0
GROUP	250	250	0.5	2.6	4.2	1.8	1.5	2.1	0.9	4.2	**99.5**
			4	1.3	38.1	1.6	34.7	2.1	19.8	4.7	**100.0**
			8	1.0	71.7	1.4	68.2	1.5	62.9	4.5	**100.0**
		500	0.5	2.8	4.3	3.4	1.2	4.4	0.6	4.9	**98.3**
			4	1.7	37.7	2.5	20.1	2.5	17.9	5.3	**100.0**
			8	1.1	72.5	1.8	61.1	1.7	61.2	5.4	**100.0**
	500	250	0.5	2.2	6.6	2.5	1.9	2.2	1.1	4.7	**100.0**
			4	0.8	60.9	1.1	65.0	1.7	50	4.1	**100.0**
			8	1.1	84.8	1.3	82.1	1.5	81.8	4.3	**100.0**
		500	0.5	2.6	7.4	3.3	1.4	3.9	0.7	5.1	**100.0**
			4	1.3	61.9	1.3	59.1	1.5	47.7	5.4	**100.0**
			8	1.4	85.6	1.8	82.0	1.8	81.7	5.6	**100.0**
TOEP–	250	250	0.5	2.8	0.6	2.9	0.5	2.2	0.4	5.2	**2.7**
			4	3.6	**67.3**	3.7	40.3	3.6	54.6	3.8	31.3
			8	3.5	**86.4**	3.3	52.7	4.0	80.0	4.0	32.8
		500	0.5	2.5	0.5	2.8	0.4	2.7	0.3	4.4	**1.7**
			4	4.3	**39.6**	4.7	27.2	3.7	25.3	4.2	28.4
			8	4.5	**50.2**	4.7	32.2	4.3	34.5	4.4	29.4
	500	250	0.5	2.3	3.5	2.7	3.2	3.0	3.1	5.1	**11.1**
			4	3.0	**93.4**	4.4	91.5	4.2	92.7	4.7	48.6
			8	3.1	**99.8**	4.0	99.6	3.6	99.7	4.3	49.8
		500	0.5	4.0	2.9	3.6	2.6	2.9	2.3	3.8	**8.8**
			4	3.6	**88.8**	3.9	59.9	4.4	86.5	5.1	46.8
			8	3.5	**98.9**	3.9	70.1	4.2	98.1	5.2	47.8

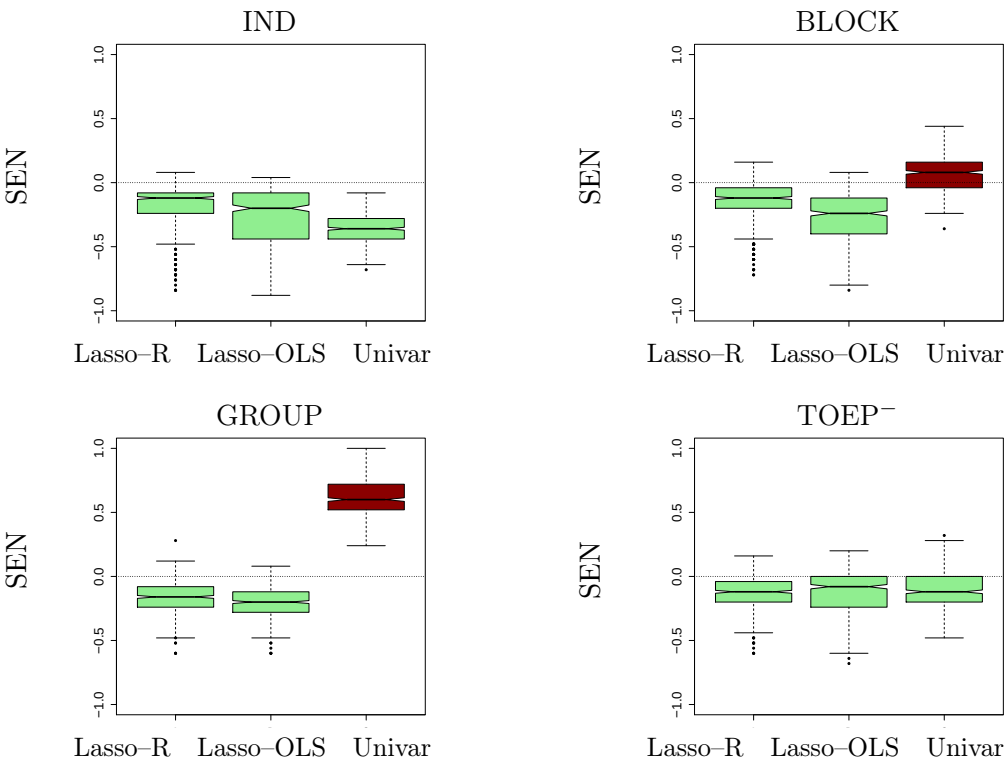


Figure 3.4 – Boxplots of sensitivity differences (the higher, the better) compared to our Lasso screening with adaptive-ridge regression cleaning: Lasso screening with ridge regression cleaning (Lasso–R), Lasso screening with OLS cleaning (Lasso–OLS), and univariate testing (Univar) , computed over the 500 simulations of each design. The tests are calibrated to control the FDR below 5%, and the boxes are light green when the FDR is actually below 5%, and dark red otherwise.

a significantly lower sensitivity, and dark red boxes indicate that the FDR is not properly controlled.

To conclude, the effect of the ridge regularization is beneficial, but the major improvements are brought by the transfer of the magnitude of weights performed by adaptive-ridge.

Comparison of our Permutation Tests versus Knockoff Test

Barber and Candes (2014) propose the knockoff procedure to provide a FDR control for the Lasso. Their method is applicable when $n > p$ and preferentially when $n > 2p$. We tested this procedure with $n = 500$ and $p = 250$. We used the `knockoff.filter` function of the `knockoff` package available in the CRAN repository (D in Figure 3.5). We also

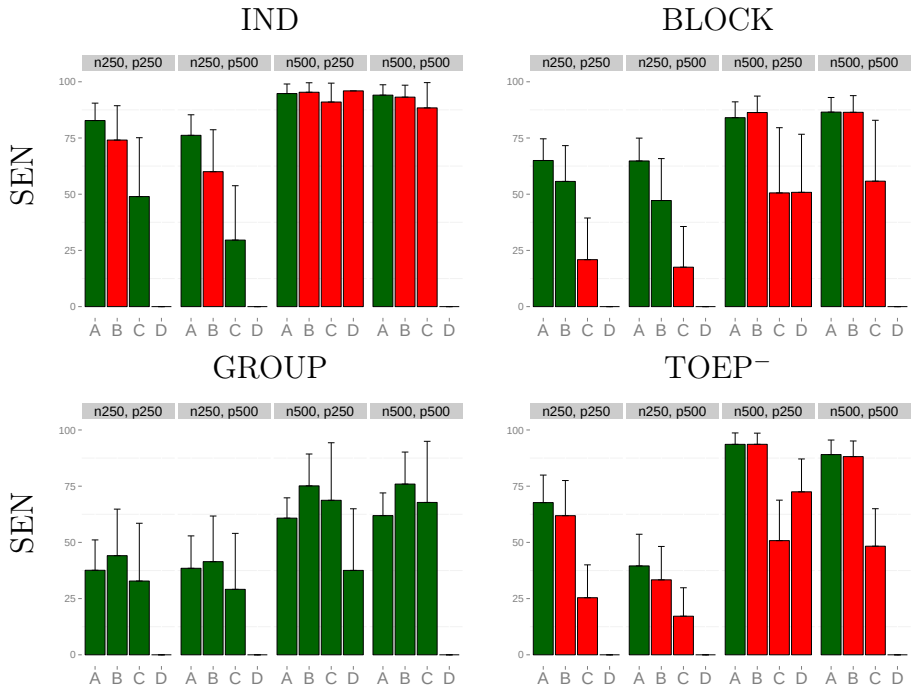


Figure 3.5 – Barplot of Sensitivity SEN for IND, BLOCK, GROUP and TOEP$^-$ designs. Bar are averaged over the 500 simulations of each design for all couples (n, p) with $n = \{250, 500\}$, $p = \{250, 500\}$ and SNR $= 4$. For each, 4 protocols are tested with Lasso screening followed by: adaptive-ridge cleaning with our statistical test (A) or with the knockoff test (B); Lasso cleaning with knockoff test (C) and when $n > p$ the Lasso with knockoff test directly on $\mathcal{D}$ without screening (D). Bars are colored in green when the FDR level is not controlled at 5% or in red instead.

tried this procedure or a part of it at the cleaning. As the Lasso [1] applied at the the screening ensures that at cleaning stage $|\hat{\mathcal{S}}(\hat{\lambda})| < n/2$. The number of available variables is smaller than the sample size of $\mathcal{D}_2$. This allows us to create the augmented matrix limited on pre-selected variables and we applied Lasso and adaptive-ridge on this. Statistics are constructed on penalties for Lasso and on coefficients for adaptive-ridge. Unfortunately, we can not succeed to applied the OLS on the augmented matrix. Indeed, the cross-product of this matrix, build with their package, is semi-definite positive so not invertible.

Figure 3.5 shows the sensitivity of the knockoff approach with a

1. The knockoff procedure does not relayon the data partition used by screen and clean

FDR level expected at 5%. Except for the GROUP design, our adaptive ridge screen and clean procedure outperformes the knockoff test. As expected the knockoff test is more sensitive for the GROUP design, where all variables in a block are relevant. Indeed, our permutation test modifies the covariance structure when the knockoff matrix conserve it. Moreover, the Fisher test measure the importance of a variable which can be masked by the others on a same block. Globally, the knockoff procedure appears to not control the FDR as the expected level even for the IND design when n is twice larger than p but for our settings the SNR is twice more harder compared to their paper.

Comparison of FDP Variability The FDR is the expected value of the FDP (see Equation (1.7)) whose variability could be an issue for the actual interpretation of results, especially in medical and biological analyses (Lin and Lee 2015). Tables 3.4 and 3.5 show the standard deviation for four variable selection procedures, when a threshold on the test statistics has to be set to control the FDR at 5%.

Table 3.4 – Standard deviation of False Discovery Proportion FDP (in %) computed over 500 simulations for each design for $n = 250$ and $p = 500$ and SNR $= 4$. Our adaptive-ridge (AR) cleaning is compared with the ridge cleaning, the original (OLS) cleaning and univariate testing (Univar). Screening is either performed by Lasso. The tests are calibrated to control the FDR below 5%, using the Benjamini-Hochberg procedure.

Simulation design	IND	BLOCK	GROUP	TOEP$^-$
AR cleaning	4.9	4.3	4.2	7.7
Ridge cleaning	8.0	6.1	8.0	8.5
OLS cleaning	9.2	7.3	8.9	9.7
Univar	6.0	2.4	6.7	7.4

When the signal is sufficient, for example with $SNR = 4$, then the FDR is controlled at the desired level but the standard deviation is higher than 5% for the major part of results. The adaptive-ridge used to clean $\hat{\mathcal{S}}(\hat{\lambda})$ performs a good control of the FDR with a lower

variability compared to Lasso + ridge or Lasso + OLS.

Table 3.5 – Standard deviation of False Discovery Proportion FDP (in %) computed over 500 simulations for each design for $n = 250$ and $p = 500$ and SNR $= 0.5$. Our adaptive-ridge (AR) cleaning is compared with the ridge cleaning, the original (OLS) cleaning and univariate testing (Univar). Screening is either performed by Lasso. The tests are calibrated to control the FDR below 5%, using the Benjamini-Hochberg procedure.

Simulation design	IND	BLOCK	GROUP	TOEP$^-$
AR cleaning	20.3	29.7	12.8	14.7
Ridge cleaning	20.4	26.8	16.3	15.6
OLS cleaning	15.9	24.7	19.5	15.1
Univar	18.0	36.7	6.3	18.9

When the noise is too high, for example with SNR $= 0.5$, the variability of FDP increases for all method and it is responsible for the absence of control the FDR in these cases. In a same time the variability of the SEN decrease dramatically because the average number of selected variables during the screening is close to zero.

The sensitivity have in a same way a strong variability. Contrary to the FDP, the variability decreases when the noise increase and the univariate approach and the adaptive-ridge seems the more stable.

Effect of the true null hypothesis proportion on FDR control For the "screen and clean" procedure, the proportion of true null hypothesis π_0 on cleaning is equal to the FDP measured after the screening and it can be estimated or known for simulations. As it mentionned previously, the FDR is controlled at a level $\pi_0\alpha$ with Benjamini-Hochberg. In the high-dimensional setup π_0 is assumed to be close to 1 but this is not realistic after the screening. For example, for SNR $= 4$ the FDR measured on $\hat{\mathcal{S}}(\hat{\lambda})$ is comprised in $[0.4, 0.8]$ and that induces a control of the FDR under the α expected level.

Figure 3.6 shows the modification on SEN and FDR when the Benjamini-Hochberg procedures take into account the effective π_0 level.

At first, the package LBE (Dalmasso et al. 2005), available in the CRAN repository, over-estimate π_0 for almost all settings but the FDR keep under the expected 5% and $\hat{\pi}_0$ is very close to π_0. As expected, the FDR is controlled at a level close to α instead lower and for all results we have a gain for the sensitivity. Unfortunately, this gain exists but is not significant instead for the TOEP$^-$ design.

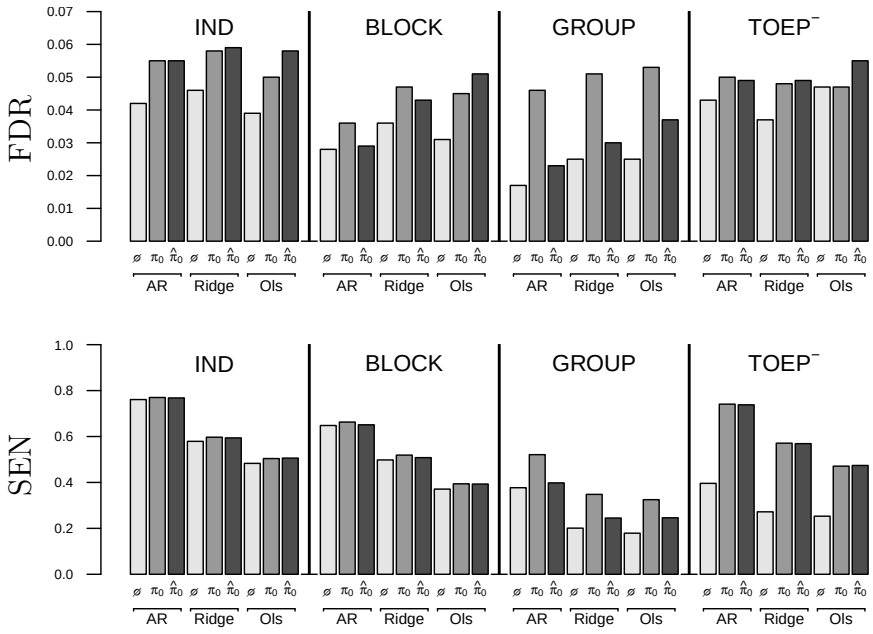


Figure 3.6 – Barplot of False Discovery Rate (FDR, top), and Sensitivity (SEN, bottom), measured when the Benjamini-Hochberg procedure takes into account the π_0 level, so where the threshold used for the i^{th} test is $\frac{i\alpha\pi_0}{m}$. Three different values of π_0 are used $\pi_0 = 1$, $\pi_0 = \text{FDR}(\hat{\mathcal{S}}(\hat{\lambda}))$, and pi_0 estimated by LBE, respectively marked by $\emptyset$ in light grey, π_0 in grey ans $\hat{\pi}_0$ in the dark gray. All results are averaged over the 500 simulations for all designs and cleaning methods for $n = 250$, $p = 500$ and SNR $= 4$.

Ordering *P*-values as an Importance Rank Some procedures exist to control the FDR on features ordered by their importance rank and not ordered by p-values (see subsection 1.1.2). These procedures may be applied easily to the Lasso and to the screening step if we rank variables on their coefficients $\hat{\boldsymbol{\beta}}$. Figure 3.7 shows the ROC curve for OLS or adaptive-ridge based on the rank of p-values and for the Lasso

based on the rank of coefficients. Except for the GROUP design, the adaptive-ridge compares favorably to the simple screening with the Lasso. Ordered p-values are more effective than ordered coefficients. The gain for feature ranking is only observed, in our simulations, when we use adaptive-ridge for cleaning and not when we use the OLS.

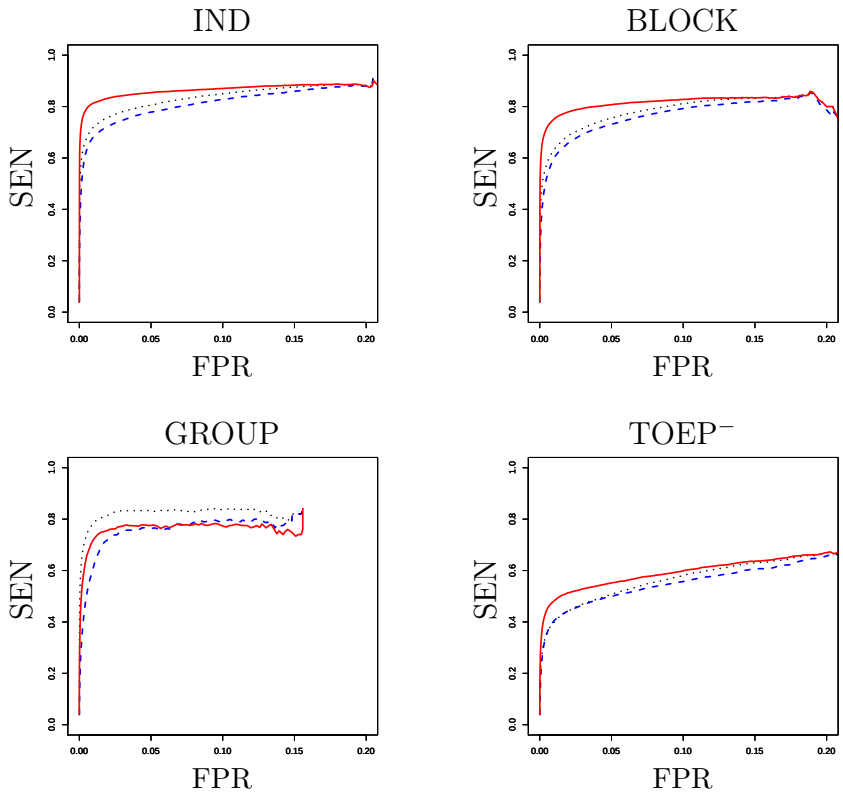


Figure 3.7 – Sensitivity SEN versus type I error rate FPR (the higher, the better). Lasso screening (black dotted line), Lasso screening followed by adaptive-ridge cleaning (red solid line) or OLS cleaning (blue dashed line). All curves are indexed by the rank of either $|\hat{\beta}_j|$ (for Lasso screening), or p-values (for Lasso and AR/OLS), and averaged over the 500 simulations of each design for $n = 250$, $p = 500$ and SNR $= 4$.

3.1.5 Selection for GWAS

We now compare the results of variable selection in a Genome Wide Association Study (GWAS) on HIV-1 infection (Dalmasso et al. 2008). One of the goal of this study was to identify genomic regions that

influence HIV-RNA levels during primary infection. Genotypes from $n = 605$ seroconverters were obtained using Illumina Sentrix Human Hap300 Beadchips. As different subregions of the major histocompatibility complex (MHC) had been shown to be associated with HIV-1 disease, the focus is set on the $p = 20,811$ Single Nucleotide Polymorphisms (SNPs) located on Chromosome 6. The $20,811$ explanatory variables are categorical variables with three levels, encoded as 1 for homozygous samples "AA", 2 for heterozygous samples "AB" and 3 for homozygous samples "BB" (where "A" and "B" correspond to the two possible alleles for each SNP). The quantitative response variable is the plasma HIV-RNA level, which is a marker of the HIV disease progression.

The Lasso screening selects $|\hat{\mathcal{S}}(\hat{\lambda})| = 20$ SNPs. Considering a 25% FDR level Dalmasso et al. (2008), the adaptive-ridge cleaning selects $|\hat{\mathcal{S}}| = 5$ SNPs as being associated with the plasma HIV-RNA, while OLS selects only $|\hat{\mathcal{S}}| = 1$ of them (see Table 3.6). Among the 12 SNPs which were identified by Dalmasso et al. (2008) from a univariate analysis in the MHC region, only 3 (rs2523619, rs214590 and rs11967684) remain selected with the proposed approach, and only one with the OLS cleaning. It is worth noting that these 12 SNPs can be clustered into two groups with high positive intra-block correlations and high negative inter-block correlations (up to $|\rho| = 0.7$). Hence, there is a high chance of confusion between these highly correlated variables. In this situation, variable selection methods working on sets of variables, such as the ones we envision in future works would be highly valuable. Those results are in line with the simulation study, in the sense that, in a similar context, the adaptive-ridge cleaning stage has a better sensitivity than OLS cleaning and is also much more conservative than univariate testing.

Table 3.6 – Adjusted p-values (in %) obtained from the Benjamini-Hochberg procedure for the five SNPs of the HIV data selected at a 25% FDR level. Our adaptive ridge (AR) cleaning is compared with the original (OLS) cleaning and with univariate testing (Univar).

SNP	Associated gene	AR cleaning	OLS cleaning	Univar
rs2523619	HLA-C*	0.0	0.1	2.10^{-5}
rs11967684	PSORS1C3/HCG27*	2.2	25.5	4.10^{-3}
rs214590	KDM1B	4.0	4.5	8.10^{-3}
rs6923486	TRAPPC3L	4.2	10.7	36.2
rs1983789	other	6.2	5.8	13.0
rs807311	LOC105377972*	14.4	9.4	0.1
rs2894181	PSORS1C3/HCG27*	15.4	30.8	2.10^{-2}
rs7749001	LOC101927314	15.6	24.8	16.7
rs4707403	LOC101928911*	22.6	30.8	5.10^{-3}
rs631175	EPHA7*	38.8	62.0	0.2
rs11155282	other	39.8	40.8	0.6
rs9459522	LOC105369172*	60.6	99.5	0.8
rs911182	MAP3K5	61.0	60.6	3.7
rs7771674	SMAP1	65.4	94.3	5.7
rs4263561	QKI*	68.2	77.9	9.4
rs2763264	LOC105378140	73.6	76.0	4.7
rs500694	LOC105374977	75.2	93.5	0.7
rs314280	LIN28B	77.2	95.8	0.5
rs236387	CPNE5	85.0	96.6	2.7
rs212388	LOC105378081	90.0	63.2	1.2

3.2 Stabilizing Screen and Clean

The screen and clean procedure of Wasserman and Roeder (2009) depends on the "screening poerty" of the subset $\hat{\mathcal{S}}(\hat{\lambda})$ chosen during the screening step. Lasso selection is very unstable as explained in Chapter 1, and we can not ensure that the true support $\mathcal{S}^\star$ belongs to $\hat{\mathcal{S}}(\hat{\lambda})$. Irrelevant covariates can be selected in place of relevant correlated covariates.

A protocol which associates "screen and clean" with "stability selection" procedures has been proposed in (Meinshausen et al. 2009). In this paper "Single-split" and "Multi-split" names were proposed to distinguish the original screen and clean method of Wasserman and Roeder (2009) and the screen and clean with resampling of Meinshausen et al. (2009). Multi-split consists to do B times the "Single-split" procedure where $\mathcal{D}_1$ and $\mathcal{D}_2$ are randomly chosen in $\mathcal{D}$ at each iteration. The main contribution of "Multi-split" is the aggregation of p-values along all the iterations for each variables. The final p-value for each variables ensure a FWER control. This protocol is summarized as follow:

1. For $b \in 1, ..., B$:
   (a) Split $\mathcal{D}$ in ${\mathcal{D}_1}^{(b)}$ and ${\mathcal{D}_2}^{(b)}$ randomly chosen where ${\mathcal{D}_1}^{(b)} \cup {\mathcal{D}_2}^{(b)} = \mathcal{D}$
   (b) Screen: Use ${\mathcal{D}_1}^{(b)}$ to estimate $\hat{\mathcal{S}}(\hat{\lambda})^{(b)}$
   (c) Clean:
   Use ${\mathcal{D}_2}^{(b)}$ to compute p-values $\tilde{P}_j^{(b)}$ by OLS for $j \in \hat{\mathcal{S}}(\hat{\lambda})^{(b)}$
   Set $\tilde{P}_j^{(b)} = 1$ for each variables $j \notin \hat{\mathcal{S}}(\hat{\lambda})^{(b)}$
   (d) Define Bonferroni-adjusted p-values as

$$P_j^{(b)} = \min\left(\tilde{P}_j^{(b)} \left|\hat{\mathcal{S}}(\hat{\lambda})^{(b)}\right|, 1\right) \tag{3.3}$$

   for $j \in 1, ..., p$

2. For $\gamma \in (0,1)$ define
$$\mathcal{Q}_j(\gamma) = \min\{1, q_\gamma(\{P_j^{(b)}/\gamma; b = 1, ..., B\})\},$$
where $q_\gamma(.)$ is the empirical γ-quantile function

3. Compute final p-values with
$$P_j = \min\left\{1, \left(1 - \log(\gamma_{min}) \inf_{\gamma \in (\gamma_{min}, 1)} \mathcal{Q}_j(\gamma)\right)\right\}, \quad (3.4)$$
where γ_{min} is the lower bound of γ

Final p-values P_j ensure a FWER control for the final selected subset $\hat{\mathcal{S}}$. Meinshausen et al. (2009) purposes an additional step to obtain a FDR control inspired by Benjamini and Hochberg (1995), as follow:

4. $\hat{\mathcal{S}}^{FDR} = \{i : i \leq h\}$,
for
$$h = \max\{i : P_{(i)} \leq \frac{i\alpha}{\sum_{i=1}^{m} i^{-1}}\}, \quad (3.5)$$
where p-values are in ascending order and $\sum_{i=1}^{p} i^{-1}$ refers to the case of general dependencies (Benjamini and Yekutieli 2001)

The number m of tested variables $|\hat{\mathcal{S}}(\hat{\lambda})|$ is not used in the definition of the rank h which is the number of selected variables $|\hat{\mathcal{S}}^{FDR}|$ under FDR control. Indeed, P_j is an aggregation of p-values already adjusted by a Bonferroni correction so, according to the author, we does not need to normalize again by $|\hat{\mathcal{S}}(\hat{\lambda})|$.

Inadequacy for FDR Control The Benjamini-Hochberg procedure can be viewed as a ranked-ponderation of a Bonferroni adjustment. Here, it is not the case because the Bonferroni adjustment is applied at each iteration and the Benjamini-Hochberg procedure only on the aggregated p-values. On a classical way the rank h (see Equation 1.9) chosen by the Benjamini-Hochberg procedure depends on $\frac{i\alpha}{m}$. Whatever the value of $i \in \{1, \ldots, m\}$, we have the threshold $\frac{i\alpha}{m} \leq \alpha$. Notice that in a classical Benjamini-Hochberg procedure a p-value upper than α will never be selected.

The specific procedure proposes by Meinshausen et al. (2009) de-

fine h (see Equation (3.5)), the number of rejected test, without taking into account the number of test m. In that case, it always exists a rank i where the threshold $i\alpha/\sum_{i=1}^{m} i^{-1}$ is upper than α. That induce a potential rejection of a test even if this test is not rejected without a FDR control procedure. For example, when $m \geq 105$ we have $i\alpha/\sum_{i=1}^{m} i^{-1} \geq 1$ for $i = m$, as it shown in figure 3.8. For this extreme case, all variables will be selected because is exists a threshold upper than 1 which is the maximal values for P_i (see Equation (3.4)).

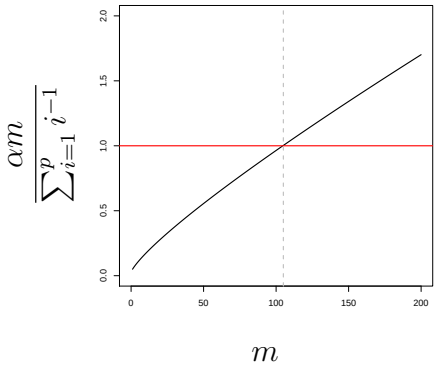


Figure 3.8 – Value of the maximum threshold calculated to define the number of rejected test h (see Equation (3.5)) when the number of variables m grows in $\{1, .., 200\}$ for a level $\alpha = 5\%$. The full red line correspond to a threshold on p-values equal to one and the dashed gray line correspond to $m = 105$ where the maximal threshold for p-values is upper than 1.

Indeed, P_i correspond to an aggregation of Bonferroni adjusted p-values. In equation (3.4) the adjusted p-values is truncated at 1, like variables which are not selected during the screening for this iteration. Due to this threshold we loose the effect of the Bonferroni adjustment and we cannot take any FDR control.

This is a big problem, but we have another mistake for this protocol. The Bonferroni adjustment is applied at each iteration and the Benjamini-Hochberg procedure only on the aggregated p-values. For the final Benjamini-Hochberg procedure we need to order variables by increasing p-values along all m tests when the Bonferroni adjustment takes into account $|\hat{\mathcal{S}}(\hat{\lambda})|$ to normalize at each iteration. So we have a

scale difference between these two procedures. In conclusion, the FDR control can not be applied with the proposition of Meinshausen et al. (2009).

3.3 Adaptive-ridge for Estimation

In sparse linear regression models, several theoretical results state conditions that ensure asymptotical support recovery, that is, the recovery of the subset of all relevant explanatory variables. One of the main result reveals a necessary and sufficient condition for the selection property of ℓ_1-regularized least squares. Several variants of this condition have been proposed, such as the irrepresentable condition, the restricted eigenvalue condition, or the mutual incoherence condition. In a nutshell, this type of condition states that the subset of truly effective variables can be retrieved exactly, provided the relevant and irrelevant covariates are not too strongly correlated. However, the rate of convergence of the Lasso may be slow and many noise variables are selected if the estimator is chosen by a predictive criterion such as cross-validation (Meinshausen 2007). These observations motivated the proposal of several two-stage procedures (Efron et al. 2004, Meinshausen 2007, Huang et al. 2008, Belloni and Chernozhukov 2013, Liu and Yu 2013). They produce models with faster convergence, smaller bias, and even, under more restrictive assumptions, oracle guarantees.

In this thesis, we experimentally investigate the large p small n designs for the Lasso+OLS (Efron et al. 2004, Belloni and Chernozhukov 2013) and Lasso+ridge (Liu and Yu 2013) procedures, comparing them to a variant based on adaptive-ridge. We do not work out the proofs of Liu and Yu (2013) to show the consistency of the adaptive-ridge variant, since we believe that this transposition would be of low interest.

3.3.1 Original Procedures

In these two-stage procedures, the support $\hat{\mathcal{S}}_\lambda$ of the sparse Lasso estimator $\hat{\boldsymbol{\beta}}(\lambda)$ of Equation (1.15) defines the set of possibly relevant variables. Then, either ordinary least squares or ridge regression is applied to the selected predictors:

$$\tilde{\boldsymbol{\beta}}(\lambda, \mu) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0, j \notin \hat{\mathcal{S}}_\lambda}{\arg\min} \; J(\boldsymbol{\beta}) + \mu \left\| \boldsymbol{\beta} \right\|_2^2 \quad ,$$

where we have the Lasso+OLS for $\mu = 0$.

Belloni and Chernozhukov (2013) and Liu and Yu (2013) work out the rates that should govern the decay of the Lasso penalty parameter λ for Lasso+OLS and Lasso+ridge respectively, but they do not propose a practical means of setting the constants so as to define the actual estimator. In their experimental section, Liu and Yu (2013) however compute λ by cross-validation, while the ridge parameter μ is set to $1/n$, thereby following the rate decay that theoretically enjoys good estimation and prediction performances.

3.3.2 Lasso+adaptive-ridge Procedure

In practice, the actual choice of the penalization parameters λ and μ is very important regarding performances. Cross-validation is commonly used to estimate the penalty parameter λ of the Lasso estimator, and we follow Liu and Yu (2013) in using this scheme for setting λ for Lasso+OLS, Lasso+ridge and Lasso+adaptive-ridge, defined as:

$$\tilde{\boldsymbol{\beta}}(\lambda, \mu) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0, j \notin \hat{\mathcal{S}}_\lambda}{\arg\min} \; J(\boldsymbol{\beta}) + \mu \sum_{j=1}^{p} \frac{\lambda}{|\hat{\beta}_j(\lambda)|} \beta_j^2 \quad ,$$

where $\hat{\beta}_j(\lambda)$ are the regression coefficients obtained by the Lasso with penalty parameter λ. Then, as setting arbitrarily $\mu = 1/n$ in Liu and Yu (2013) can lead to very bad performances for Lasso+ridge or Lasso+adaptive-ridge, we also chose to set μ by cross-validation.

Note that, if applied naively, this serial selection process is prone to overfitting, in the sense that the variables selected by the Lasso are

likely to be correlated with the response variable, resulting in optimistic conclusions regarding variable importance, a phenomenon known as Freedman's paradox in model selection (see Freedman 1983). Our protocol consists in cross-validating the complete serial process to select μ once λ has been chosen in the screening stage of the procedure (that is, λ is fixed, but $\hat{\mathcal{S}}_\lambda$ is recomputed at each fold of the cross-validation process). Finally, following Meinshausen (2007), we set jointly λ and μ by cross-validation, so that the λ parameter of the Lasso screening is optimized so as to minimize the expected prediction error of the Lasso+adaptive-ridge estimator instead of the error of the Lasso estimator itself.

3.3.3 Results for Estimation

In the following, we discuss the IND, BLOCK, GROUP and TOEP$^-$ designs with $n = 250$, $p = 500$, $|\mathcal{S}^\star| = 50$, $\rho = 0.5$ and a size block equal to 25. We report results with three different noise levels. The relative behavior of the estimation methods is similar for high and medium noise levels (respectively SNR $= 4$ and SNR $= 8$), with more significant differences for medium noise levels. The situation then drastically changes for the low noise level (SNR $= 32$), where two stage methods become beneficial.

We compare the variants of the two-stage estimation methods based on the predictive mean squared error. Similar conclusions would be drawn from the accuracy measures on the vector of coefficients $\boldsymbol{\beta}^\star$. Figure 3.9 displays the boxplots of prediction error obtained over 500 simulations for each design.

There is no benefit in a post-Lasso estimation step for high and medium noise levels (SNR $\in \{4, 8\}$). OLS and ridge post-processing then have important detrimental effects and adaptive-ridge has still a slight unfavorable effect. It is only when the two-step procedure is jointly optimized with respect to the two penalization parameters (by cross-validation), that some improvements become visible for the first

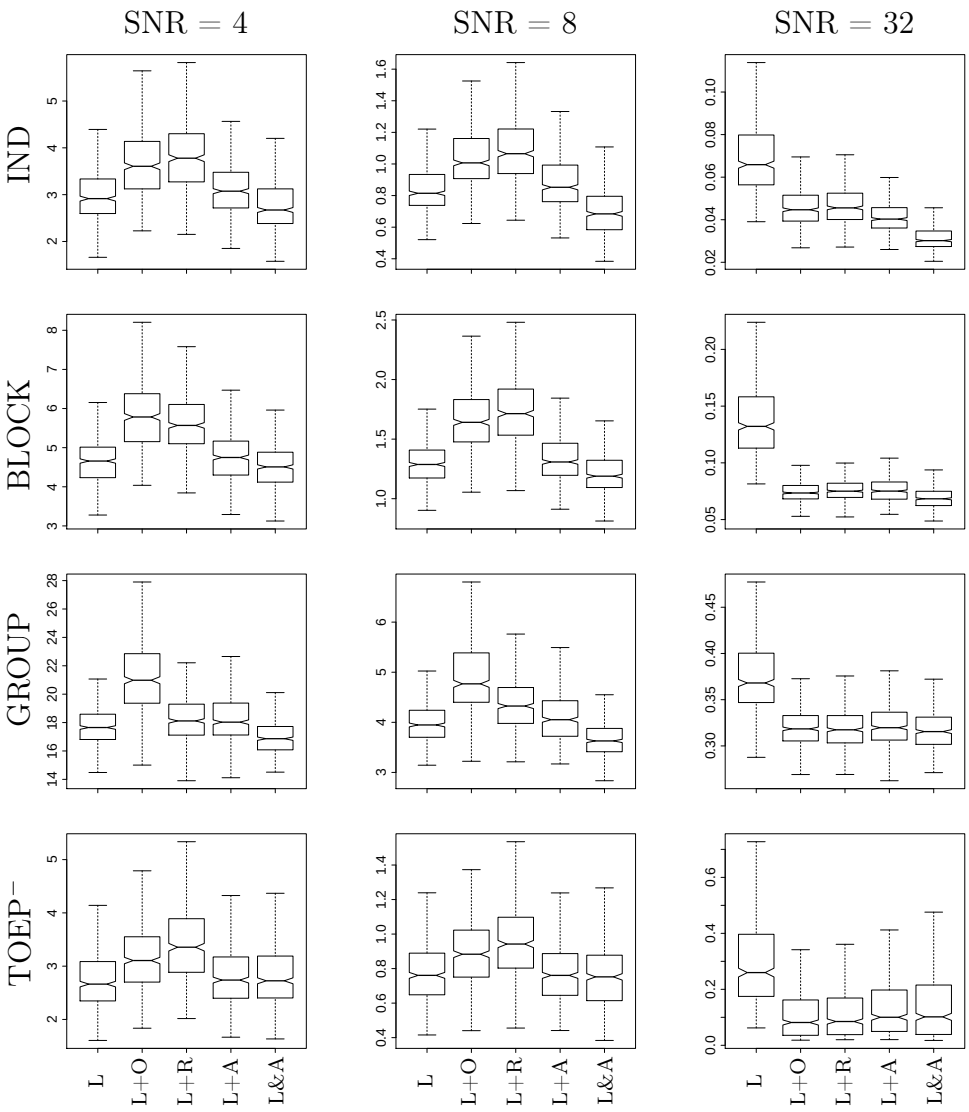


Figure 3.9 – Mean prediction error computed over 500 simulations for each design. Lasso direct estimation (L) is compared to: Lasso screening followed by OLS estimation (L+O), Lasso screening followed by ridge estimation (L+R), Lasso screening followed by adaptive-ridge estimation (L+A), jointly optimized Lasso screening with adaptive-ridge estimation (L&A).

three setups.

When the signal-to-noise ratio is high (SNR = 32), Lasso highly benefits from the second stage whatever it may be (OLS, ridge or adaptive-ridge). There is a slight edge to adaptive-ridge when variables are independent, but otherwise all methods are at par. Globally, the best option here consists again in jointly optimizing the two stages

with respect to the two penalization parameters; some additional improvements come into view.

Compared to previous studies, which mainly focused on large sample and/or low-noise settings, our experiments demonstrate that post-Lasso estimation can have consequential beneficial or detrimental effects in small sample regression. In addition to the experimental design, the results vary also considerably according to the strategy governing the choice of the penalty parameters. Other experiments (not shown here) attest that using more stringent screening stages (using the so-called "1-SE rule" of Breiman et al. 1984, that chooses the highest penalty within one standard deviation of the minimum of cross-validation) lead to better post-Lasso estimation in some experimental setups, but this is not systematic: in the TOEP$^-$ design, this is by far the least favorable option. Overall, the joint optimization with respect to the two penalization parameters seems to be a very challenging contender. This is also true when the Lasso screening is followed by OLS or ridge regression. The joint optimization of penalization parameter favors a stringent Lasso screening compared to the strategy based on serial cross-validation, and a less stringent one compared to the 1-SE rule. Though this solution is the most expensive from the computational viewpoint, it seems to be also the most effective one regarding predictive mean squared error.

3.4 Discussion

We propose to use the magnitude of regression coefficients in two-stage variable selection procedures. First, we use the connection between the Lasso and adaptive-ridge (Grandvalet 1998) to convey more information from the screening stage to the second stage: the magnitude of the coefficients estimated at the screening stage is transferred to the second stage through penalty parameters.

On the theoretical side, we would like to back-up the empirical improvements that have been almost systematically observed by an

apposite analysis. Our preliminary results in the orthonormal setting (which are precise and accurate up to numerical integration errors) lack the clarity of purely analytical results. Besides, the orthonormal setting is not adapted to the high-dimensional setting.

Empirically, our procedure brings marginal improvements when the second stage aims at improving the regression coefficients (Belloni and Chernozhukov 2013, Liu and Yu 2013), and it provides sensible improvements compared to the original screen and clean procedure (Wasserman and Roeder 2009) when assessing the uncertainties pertaining to the selection of relevant variables. In the first setup, screening and estimation are performed on the same data set, whereas in the second one, the screening and cleaning stages operate on two distinct subsamples of data: the transfer is more valuable in this situation.

Regarding post-Lasso estimation, our experiments demonstrate that two-stage methods can have consequential beneficial or detrimental effects in small sample regression. The results vary considerably according to the strategy governing the choice of the penalty parameters, but the joint optimization with respect to the two penalization parameters is the most effective one regarding predictive mean squared error.

For screen and clean, we obtained a better control of the False Discovery Rate, which extends to more difficult settings, with high correlations between variables. Furthermore, the sensitivity obtained by our cleaning stage is always as good, and often much better than the one based on the ordinary least squares. The penalized second step also allows for a less severe screening, since the second stage can then handle more than $n/2$ variables. Our procedure can thus be employed in very high-dimensional settings, as the screening property (that is, in the words of Bühlmann (2013), the ability of the Lasso to select all relevant variables) is more easily fulfilled, which is essential for a reliable control of the false discovery rate.

Several interesting directions are left for future works. The second stage can accommodate arbitrary penalties, and our efficient implementation applies to all penalties for which a quadratic variational

formulation can be derived. This is particularly appealing for structured penalties such as the fused-lasso or the group-Lasso, allowing to use the knowledge of groups at the second stage, through penalization coefficients. This extension is the subject of following chapter, where we concentrate on the inference problem. Instead of testing the significance of the individual contribution of variables, which seems to be rather weak in our application domain, we will target questions related to the significance of groups of variables. The loss of precision in the question is expected to bring some confidence in the answer.

4. Two-Stage Selection of Groups of Variables

The previous chapter presented our contributions to "flat" variable selection without any structuration of variables. However, in many situations the variables are organized in groups that are expected to be globally relevant or not. For example, in GWAS studies (see 3.1.5) the real task is to retrieve genomic regions associated to a biological response. Then, SNPs are representative of biological groups. By example, the Major Histocompatibility Complex (MHC) is a sub-region of chromosome 6 that can be considered as a group of interest. Besides their biological signification, groups of SNPs also make sense from a statistical viewpoint since they comprise highly correlated variables. It is therefore difficult to differentiate among individual contributions, and it may be regarded as more apposite to consider groups instead of variables thereby hopefully trading correctness for accuracy.

This chapter presents the our proposal for a group-wise screen and clean procedure. Screening is then based on Lasso variants that take into account groups of variables. Cleaning is still based on adaptive-ridge, which now incorporates a measure of the importance of groups of variables, to conclude on the statistical significance of these groups.

4.1 Screen and Clean for Group Selection

Our adaptation of the screen and clean procedure to the selection of groups is based on the variants of Lasso that have been proposed to

Table 4.1 – Lasso, elastic-net (E.-net), group-Lasso (GL), sparse-group-Lasso (SGL), and Cluster Representative Lasso (CRL) penalties with their adaptive-ridge counterpart. The adaptive-ridge penalty term is expressed as a function of the solution to the corresponding original problem.

	Original penalty	Adaptive-ridge penalty
Lasso	$\lambda_1 \|\boldsymbol{\beta}\|_1$	$\sum_{j=1}^{p} \frac{\lambda_1}{\|\hat{\beta}_j\|} \beta_j^2$
E.-net	$\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$	$\sum_{j=1}^{p} \left(\frac{\lambda_1}{\|\hat{\beta}_j\|} + \lambda_2 \right) \beta_j^2$
GL	$\lambda_2 \sum_{g=1}^{G} \sqrt{w_g} \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2$	$\sum_{g=1}^{G} \sum_{j \in \mathcal{G}_g} \frac{\lambda_2 \sqrt{w_g}}{\left\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_g}\right\|_2} \beta_j^2$
SGL	$\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{g=1}^{G} \sqrt{w_g} \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2$	$\sum_{g=1}^{G} \sum_{j \in \mathcal{G}_g} \left(\frac{\lambda_1}{\|\hat{\beta}_j\|} + \frac{\lambda_2 \sqrt{w_g}}{\left\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_g}\right\|_2} \right) \beta_j^2$
CRL	$\lambda_1 \|\boldsymbol{\beta}\|_1$	$\sum_{g=1}^{G} \frac{\lambda_1}{\|\hat{\beta}_g\|} \beta_g^2$

deal with disjoint groups of variables forming a partition. These three methods, introduced in Chapter 1, are group-Lasso (GL), sparse-group-Lasso (SGL) and Cluster Representative Lasso (CRL), which consists in applying Lasso on a $n \times G$ pseudo design matrix, where each group is represented by a unique variable formed by aggregating the variables of this group.

4.1.1 Adaptive-ridge for Group Penalties

Table 4.1 lists the main variants of Lasso of interest here, together with their adaptive-ridge counterpart. The G groups of variables are indexed by $\mathcal{G}_g,\ g = 1, \ldots, G$ of size w_g, except for CRL, where there is a single variable per group in the pseudo design matrix.

We detail below the derivation of he adaptive-ridge counterpart of sparse-group-Lasso, which follows the steps already used for elastic-net in Chapter 3. We start from a quadratic variational formulation of

sparse-group-Lasso:

$$\begin{aligned}\min_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\tau}\in\mathbb{R}^p,\boldsymbol{\gamma}\in\mathbb{R}^G} \; & J(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} \frac{1}{\tau_j}\beta_j^2 + \lambda_2 \sum_{g=1}^{G} \frac{\omega_g}{\gamma_g} \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2^2 \\ \text{s. t. } & \sum_{j=1}^{p} \tau_j - \|\boldsymbol{\beta}\|_1 \leq 0 \\ & \sum_{g=1}^{G} \gamma_g - \sum_{g=1}^{G} \sqrt{\omega_g} \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2 \leq 0 \\ & \tau_j \geq 0 \;\;,\;\; j = 1,\ldots,p \\ & \gamma_g \geq 0 \;\;,\;\; g = 1,\ldots,G \;\;. \end{aligned} \tag{4.1}$$

The variables $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$ introduced in this formulation, which adapt the ℓ_2 penalties to the data, can be shown to lead to the following adaptive-ridge penalty:

$$\sum_{j=1}^{p} \beta_j^2 \left(\frac{\lambda_1}{|\hat{\beta}_j(\lambda_1,\lambda_2)|} + \frac{\lambda_2\sqrt{w_g}}{\left\|\hat{\boldsymbol{\beta}}_{\mathcal{G}_g}(\lambda_1,\lambda_2)\right\|_2} \right) \;\;, \tag{4.2}$$

where the coefficients $\hat{\beta}_j(\lambda_1,\lambda_2)$ are the solution to the sparse-group-Lasso problem (1.20). Using this adaptive-ℓ_2 penalty returns the original sparse-group-Lasso estimator, as shown in the following lemma.

Lemma 2. *The quadratic penalty in $\boldsymbol{\beta}$ in* (4.1) *acts as the sparse-group-Lasso penalty $\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{g=1}^{G} \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2$.*

Proof. The Lagrangian of Problem (4.1) is:

$$\begin{aligned} L(\boldsymbol{\beta}) =& J(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} \frac{1}{\tau_j}\beta_j^2 + \lambda_2 \sum_{g=1}^{G} \frac{\omega_g}{\gamma_g} \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2^2 \\ & + \nu_0 \left(\sum_{j=1}^{p} \tau_j - \|\boldsymbol{\beta}\|_1 \right) - \sum_{j=1}^{p} \nu_j \tau_j \\ & + \eta_0 \left(\sum_{g=1}^{G} \gamma_g - \sum_{g=1}^{G} \sqrt{\omega_g} \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2 \right) - \sum_{g=1}^{G} \eta_g \gamma_g \;\;. \end{aligned}$$

Thus, the first order optimality conditions for τ_j are

$$\begin{aligned} \frac{\partial L}{\partial \tau_j}(\tau_j^\star) = 0 &\Leftrightarrow -\frac{\lambda_1 \beta_j^2}{\tau_j^{\star 2}} + \nu_0 - \nu_j = 0 \\ &\Leftrightarrow -\lambda_1 \beta_j^2 + \nu_0\, \tau_j^{\star 2} - \nu_j\, \tau_j^{\star 2} = 0 \\ &\Rightarrow -\lambda_1 \beta_j^2 + \nu_0\, \tau_j^{\star 2} = 0 \;\;, \end{aligned}$$

where the term in ν_j vanishes due to complementary slackness, which implies here $\nu_j \tau_j^\star = 0$. With the constraints of Problem (4.1), this equation implies $\tau_j^\star = |\beta_j|$.

Likely, the first order optimality conditions for γ_g are

$$\begin{aligned}\frac{\partial L}{\partial \gamma_g}(\gamma_g^\star) = 0 &\Leftrightarrow -\frac{\lambda_2 \omega_g \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2^2}{\gamma_g^{\star 2}} + \eta_0 - \eta_g = 0 \\ &\Leftrightarrow -\lambda_2 \omega_g \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2^2 + \eta_0 \, \gamma_g^{\star 2} - \eta_g \, \gamma_g^{\star 2} = 0 \\ &\Rightarrow -\lambda_2 \omega_g \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2^2 + \eta_0 \, \gamma_g^{\star 2} = 0 \ ,\end{aligned}$$

where the term in η_g vanishes due to complementary slackness, which implies here $\eta_g \gamma_g^\star = 0$. With the constraints of Problem (4.1), this equation implies $\gamma_g^\star = \sqrt{\omega_g} \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2$.

Hence, Problem (4.1) is equivalent to:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} J(\boldsymbol{\beta}) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{g=1}^{G} \sqrt{w_g} \left\|\boldsymbol{\beta}_{\mathcal{G}_g}\right\|_2 \ ,$$

which is the original sparse-group-Lasso formulation. □

4.1.2 Hypothesis Testing For Group Selection

The Fisher statistic (1.5) measures the normalized difference between the residuals of the full model Ω and the reduced model ω. Up to now, we used this statistic to test the significance of a single variable, by defining the reduced model as the full model deprived from a single variable. In the context of group selection, the reduced model is deprived from a whole group of variables. Otherwise, the statistic remains essentially the same, except for the change in degrees of freedom that arises from the change in group sizes. It measures the overall significance of the variables of the group and thereby the significance of the group.

Note that the efficient implementation of the permutation F-test given in Chapter 2 holds for groups. As detailed in the calculations, the accelerations tricks for the permutation of one variable apply without troubles to arbitrary group sizes.

4.1.3 Procedure for Group Selection

The screen and clean procedure requires little changes to accommodate the selection of predefined groups of variables. Regarding screening, groups enter the process through the group penalties presented in Section 4.1.1. Variable importance is then transferred to the cleaning stage through the apposite adaptive-ridge penalty. Finally, the FDR control (at the group level) is controlled at the cleaning step using the permutation F-test for groups of variables.

The complete procedure for sparse-group-Lasso (and thus for group-Lasso) screening is summarized as follows:

Input: dataset $\mathcal{D}$ comprising the $n \times p$ covariate matrix $\mathbf{X}$ and the n-dimensional response vector $\mathbf{y}$, group indices $\{\mathcal{G}_g\}_{g=1}^{G}$, trial values for the $l1$ penalty Λ_1, trial values for the $l2$ penalty Λ_2, and B the number of permutations

Output: index $\hat{\mathcal{S}}$ of groups of variables selected during screening, and P a vector with p-values associated to these groups

0. Split randomly $\mathcal{D}$ in two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$ of size $n/2$
1. "Screening" on $\mathcal{D}_1$:
   (a) Select $(\hat{\lambda}_1, \hat{\lambda}_2)$ in $\Lambda_1 \times \Lambda_2$ by 10-fold cross-validation
   (b) Define $\hat{\mathcal{S}} = \{g \in \{1, \dots, G\} : \left\| \hat{\boldsymbol{\beta}}_{\mathcal{G}_g}(\hat{\lambda}_1, \hat{\lambda}_2) \right\|_2 \neq 0\}$
2. "Cleaning" on $\mathcal{D}_2$:
   (a) Compute AR estimate $\hat{\boldsymbol{\beta}}_{\{\mathcal{G}_g\}_{g\in\hat{\mathcal{S}}}}(\hat{\lambda}_1, \hat{\lambda}_2)$ from $\hat{\boldsymbol{\beta}}_{\{\mathcal{G}_g\}_{g\in\hat{\mathcal{S}}}}(\hat{\lambda}_1, \hat{\lambda}_2)$
   (b) Compute RSS^{Ω}, the residual sum of squares for $\hat{\boldsymbol{\beta}}_{\{\mathcal{G}_g\}_{g\in\hat{\mathcal{S}}}}$ on $\mathcal{D}_2$
   (c) for $g \in \hat{\mathcal{S}}$:
   - Compute AR estimate $\hat{\boldsymbol{\beta}}_{\{\mathcal{G}_k\}_{k\in\hat{\mathcal{S}}\setminus\{g\}}}(\hat{\lambda}_1, \hat{\lambda}_2)$
   - Compute RSS^{ω} for $\hat{\boldsymbol{\beta}}_{\{\mathcal{G}_k\}_{k\in\hat{\mathcal{S}}\setminus\{g\}}}$ on $\mathcal{D}_2$
   - $F = \frac{RSS^{\omega} - RSS^{\Omega}}{RSS^{\Omega}}$
   - for $b \in \{1, \dots, B\}$:
     — $\mathbf{X}^{(b)}_{\mathcal{D}_2 \cdot} = \mathbf{X}_{\mathcal{D}_2 \cdot}$.
     — Permute randomly the lines of $\mathbf{X}^{(b)}_{\mathcal{D}_2 \mathcal{G}_g}$
     — Compute AR estimate $\hat{\boldsymbol{\beta}}^{(b)}_{\{\mathcal{G}_g\}_{g\in\hat{\mathcal{S}}}}(\hat{\lambda}_1, \hat{\lambda}_2)$ on $\mathbf{X}^{(b)}_{\mathcal{D}_2 \cdot}$.
     — Compute $RSS^{\Omega(b)}$ for $\hat{\boldsymbol{\beta}}^{(b)}_{\{\mathcal{G}_g\}_{g\in\hat{\mathcal{S}}}}$
     — $F^{(b)} = \frac{RSS^{\omega} - RSS^{\Omega(b)}}{RSS^{\Omega(b)}}$
   - Compute the empirical p-value $P_g = \frac{1}{B}\#\{F \leq F^{(b)}\}$

4.2 Numerical Experiments

4.2.1 Simulation Models

In the following, we elaborate on the BLOCK design presented in Chapter 2, with $n \in \{125, 250\}$, 25 relevant variables among $p = 500$ ones, $\rho \in \{0.2, 0.5, 0.8\}$ and SNR $= 4$. We increased here the number of blocks from 20 to 50 to have a more challenging group-wise selection problem (the block size, originally set to 25, goes down accordingly to 10), and the number of relevant blocks is set to 10, meaning that each relevant block contains an average of 2.5 relevant variables. The different levels of within-bloc correlation adjust the specificity of relevant variables.

On these designs, we apply the group-wise selection methods using three different definitions of groups:

COR 50 groups corresponding to the 50 blocks of correlated variables;

REL same as COR, except that the relevant variables belong to an extra group;

ALEA 50 groups of size 10 are randomly picked.

These three definitions allow for testing the usual situation where groups comply with correlations (COR), a favorable situation where a single group gathers the relevant variables, but where correlation may act as a distractor (REL), and an unfavorable case where groups are not related to relevance, and where correlation may furthermore act as a distractor (ALEA).

These definitions, schematized in the figure 4.1, allow to test if our procedures are more apposite for groupings agreeing with correlations or only to the signal of relevant features.

4.2.2 Results

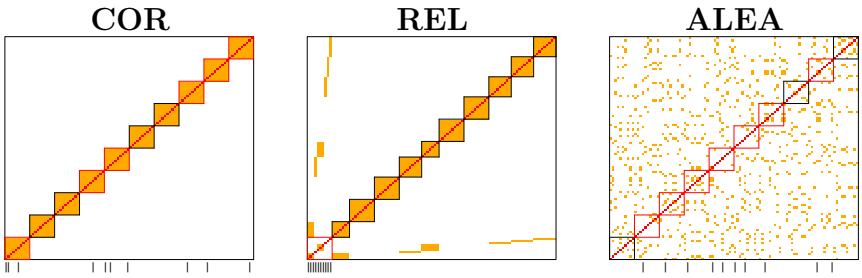


Figure 4.1 – Schematic representation of groups definition : COR (left), REL (center) and ALEA (right). Degrees of absolute correlation variyng from the white for weak correlation and to rhe read for strong correlation. Box represents how groups are indexed and a red box represent a groups considered as relevant and a black box instead. Ticks under the X-axis represent which variables have a non-null coefficient in $\boldsymbol{\beta}$.

Our group-wise procedures produce p-values for groups. They are either derived from a permutation test for AR cleaning or from an exact Fisher test for OLS cleaning. We first assess the calibration of this permutation test.

Group-Wise Permutation Test We focus our attention on the FPR and the power of the tests to assess the relevance and the performance of our permutation tests. We recall that a group is relevant provided it contains at least one relevant variable. The power measures the capacity to reject $\mathcal{H}_0$ for relevant groups, which is different to the global sensitivity of the procedure. Indeed, the global sensitivity show our capacity to retrieve the relevant variables passing the screening stage, and thereby also relates to this screening stage.

Table 4.2 displays the FPR and power for adaptive-ridge cleaning. The FPR level is controlled close to 5% as expected, except for REL grouping, and especially for the procedure based on CRL. When CRL is used at screening, all variables in a group are combined and this is likely to harm the correlation with the response variable in the REL grouping. In this case, all relevant signal disappears and it is replaced by (correlated) variables of others groups. This phenomenon explains

Table 4.2 – Type I error (FPR) and power (POW) for group-Lasso (GL), sparse-group-Lasso (SGL) and Cluster Representative Lasso (CRL) screening followed by adaptive-ridge cleaning. The statistics are computed over the 500 simulations with $p = 500$ for COR, REL and ALEA groupings. The tests are calibrated to control the FPR at 5%. No results are reported for CRL for REL grouping where no relevant variable ever passed the screening stage.

	n	ρ	COR		REL		ALEA	
			FPR	POW	FPR	POW	FPR	POW
GL	125	0.2	5.3	53.4	4.6	100.0	4.2	18.0
		0.5	5.1	62.7	4.8	100.0	2.1	13.3
		0.8	4.8	76.4	7.4	96.8	1.3	8.5
	250	0.2	5.1	90.5	4.8	100.0	3.3	47.1
		0.5	4.7	91.1	4.6	100.0	0.7	32.4
		0.8	4.9	89.3	5.1	100.0	0.3	14.7
SGL	125	0.2	4.8	63.9	4.4	100.0	4.9	33.8
		0.5	4.6	70.7	4.9	100.0	4.2	35.3
		0.8	5.3	78.7	6.6	99.3	5.5	31.8
	250	0.2	4.8	96.3	4.8	100.0	4.7	84.1
		0.5	4.8	95.3	4.8	100.0	3.8	77.9
		0.8	5.6	92.7	4.7	100.0	4.0	60.2
CRL	125	0.2	4.6	34.4	9.3	–	4.2	10.3
		0.5	5.1	56.4	19.9	–	7.0	14.9
		0.8	5.5	73.2	27.8	–	11.1	19.8
	250	0.2	4.9	54.2	14.9	–	5.5	16.1
		0.5	4.7	70.5	28.1	–	10.1	24.5
		0.8	5.0	84.9	33.6	–	18.9	30.4

the growth of the FPR for CRL-REL when ρ is high. The GL and SGL procedures easily deal with REL because all signal is gathered in a unique block which notably differs from the other blocks of irrelevant variables, even with high correlation. Of course, this extreme setup is not realist, but it checks the capacity of procedures to detect blocks that contain signal without mistaking them with correlated blocks.

The ALEA grouping lead to similar conclusions, but the sensitivity decreases dramatically due to the inappropriate group definition.

In COR, grouping agrees with correlation. Then, FPR is controlled with SGL having a highr power than the GL approach. That could be explaining by a removal of irrelevant variables on a group containing relevant variables. Correlation levels have no significant effect on the power of tests for GL and SGL screening when n is sufficiently large.

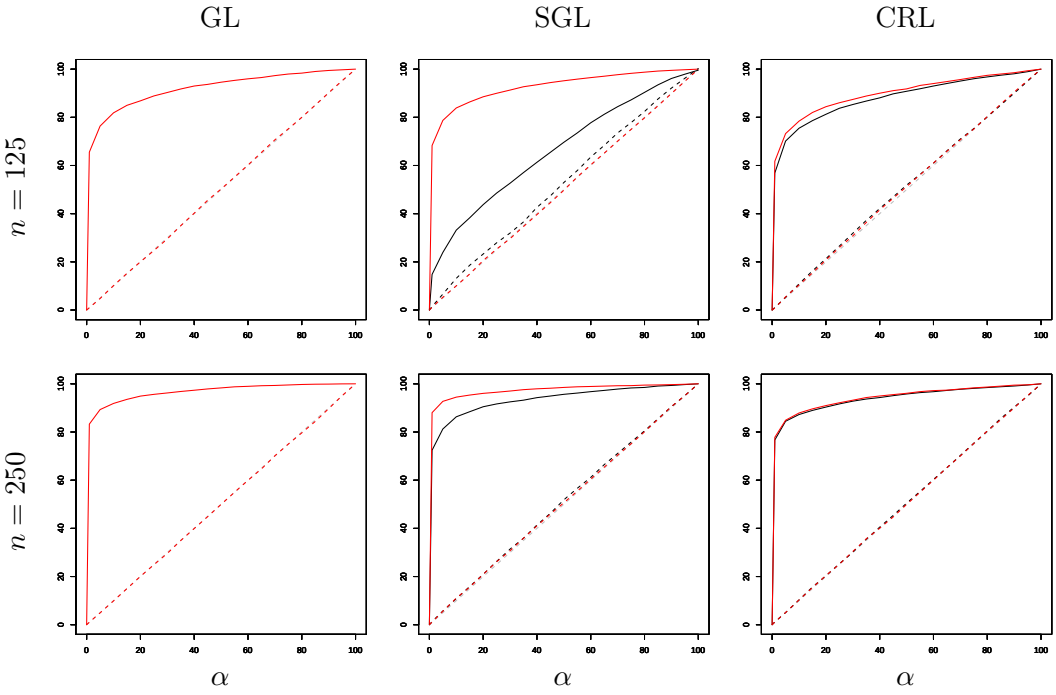


Figure 4.2 – POWER (solid lines) and FPR (dashed lines) measures when group-Lasso (GL, at left), sparse-group-Lasso (SGL, at middle) and Cluster Representative Lasso (CRL, at right) are followed by adaptive-ridge (AR, in red) or Ordinary Least Square regression (OLS, in black). Curves for GL followed by OLS is not shown because only 30 simulations over the 500 allows to use OLS at screening. The gray dashed line corresponds to the expected FPR level. All measures are estimated for different α corresponding to a threshold on p-values (x-axis) and all axes are expressed in percent. The statistics are computed over the 500 simulations with $p = 500$ and $\rho = 0.8$.

Figure 4.2 show a perfect control of the Type I error for the OLS with an exact F-test or for the adaptive-ridge with our permutation tests. That confirms the pertinence or the adaptive-ridge with the same error control without sample size constraint and with a significant better power.

Performance of the Overall Procedure We now assess the overall screen and clean performance based on the group-wise false discovery rate and sensitivity. As before, the control of the false discovery rate is set by a Benjamini-Hochberg procedure.

Table 4.3 displays the overall sensitivity of our procedure for FDR controlled at the 5% level at the cleaning stage. Without cleaning, the false discovery rate is about 70% in average, whereas after cleaning, the actual false discovery rate is at 2% in average, and below the 5% control level (not shown in table). This control comes at a price in sensitivity, but AR cleaning clearly outperforms OLS cleaning in this respect, especially with GL and SGL screening. Figure 4.3 shows that this benefit does not impact the actual FDR levels, which are roughly identical for both cleaning variants.

Table 4.3 – Sensitivity (SEN) of our screen and clean procedure with group-Lasso (GL), sparse-group-Lasso (SGL) or Cluster Representative Lasso (CRL) screening, without cleaning (W/o), or followed by adaptive-ridge (AR) or Ordinary Least Squares (OLS) cleaning. The statistics are computed over 500 simulations, with $p = 500$ and the COR grouping. Cleaning is calibrated to control the FDR at 5%.

n	ρ	GL			SGL			CRL		
		W/o	OLS	AR	W/o	OLS	AR	W/o	OLS	AR
125	0.2	82.3	0.2	26.7	84.8	1.6	40.1	43.9	6.9	8.3
	0.5	83.5	0.0	38.2	85.9	2.1	49.1	62.4	22.4	26.3
	0.8	83.5	0.1	56.6	86.4	5.9	60.9	80.6	46.8	51.7
250	0.2	97.1	0.0	83.3	98.5	31.4	93.8	52.6	20.0	20.9
	0.5	97.0	0.0	84.3	98.4	35.9	92.4	77.6	45.1	47.0
	0.8	93.1	0.4	79.6	95.4	50.3	85.4	88.1	69.5	70.3

The difference in sensitivity between OLS and AR cleaning is somewhat inflated in Table 4.3, in the sense that, for OLS, FDR control is enforced by setting the sensitivity to zero when the OLS estimator is not defined, that is, when the number of variables passing the screening set is greater than the cleaning sample size $n/2$.

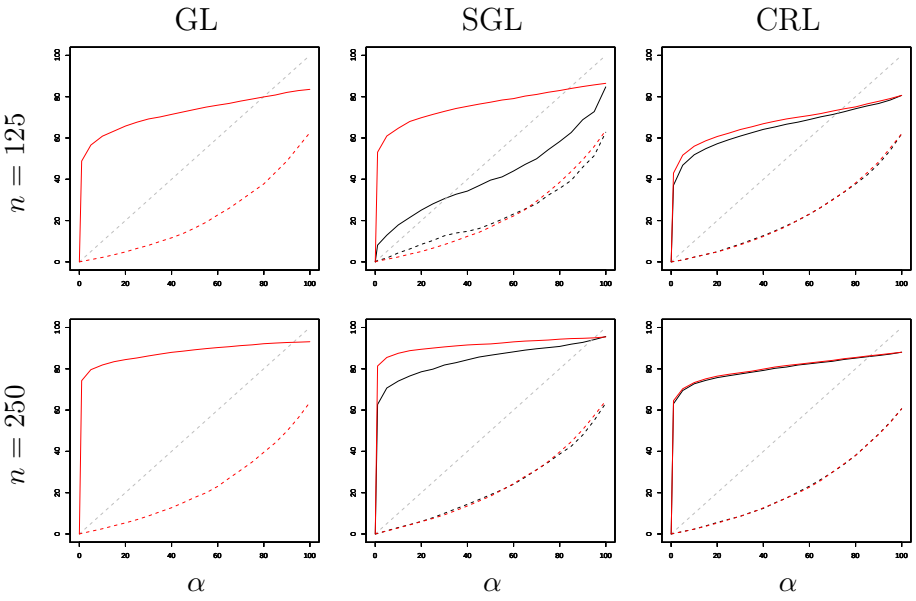


Figure 4.3 – Sensitivity (solid) and False Discovery Rate versus control level α when group-Lasso (GL, left), sparse-group-Lasso (SGL, center) and Cluster Representative Lasso (CRL, right) screening are followed by cleaning with adaptive-ridge (red) and Ordinary Least Square regression (black). The statistics are averaged over the 500 simulations with $p = 500$ and $\rho = 0.8$. The gray dashed diagonals represent the assumed FDR level. Results for GL followed by OLS are not shown because only 30 simulations out of 500 are amenable to OLS cleaning.

Figure 4.3 provides a more complete view of the merits of OLS and AR cleaning stages when OLS is well-defined. For all screening variants and all levels of control of FDR, AR cleaning clearly outperforms OLS cleaning. Finally, Figure 4.4 illustrates that the differences in sensitivity are highly variable, though significantly in favor of AR cleaning.

4.3 Selection for GWAS

We now analyze the HIV dataset introduced in Section 3.1.5, using the tools for group-wise analysis introduced above. SNPs are known to be highly correlated due to their proximity along chromosomes and to the co-regulation of their corresponding genes. The individual statistical effects of SNPs should thus be rather incremental, and a group analysis may be more appropriate from a biological viewpoint, in order

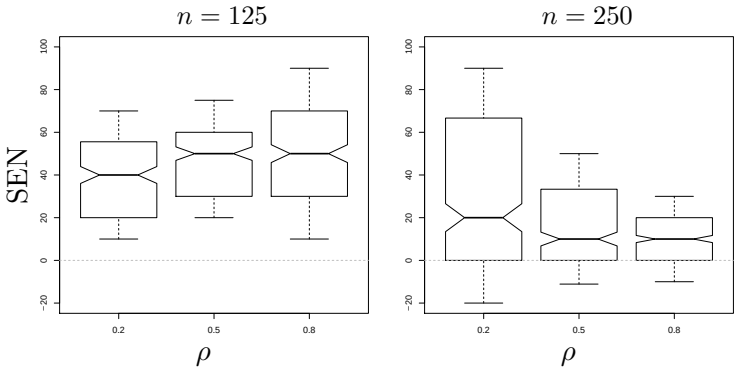


Figure 4.4 – Boxplots of differences in sensitivity between adaptive-ridge and OLS cleaning after sparse-group-Lasso cleaning (the upper, the better for AR cleaning). Each boxplot summarizes 500 simulations where the procedure is calibrated to control the FDR below 5% for each design/grouping.

to retrieve genes or regions associated with the HIV viral load.

The definition of relevant groups of SNPs is however not straightforward. SNPs are markers positioned at regular intervals along the genome, either in genic or intergenic regions. Half of the 20,811 SNPs are located in intergenic regions, and these regions contain an important part of the material responsible for gene regulation. Hence, even though the membership of SNPs to gene is readily available in public databases, clustering SNPs based only on gene membership does not appear to be judicious.

Here, we follow Dehman et al. (2015), who proposed to use the linkage disequilibrium between SNPs to define clusters. Linkage disequilibrium quantifies the preferential association between pairs of SNPs. Knowing the proportion of allele on each SNP, we can estimate the expected proportion of allele association between two SNPs. Linkage disequilibrium is defined as the difference between the expected proportion and the observed one.

Dehman et al. (2015) used this measure in a hierarchical classification based on Ward's criterion, thereby clustering the 20,811 SNPs in 1,756 groups, using the dataset presented here. These groups mainly represent sets of genes with their associated intergenic regions. Figure

4.5 displays a histogram of group sizes, where we observe that most groups contain less than 20 SNPs.

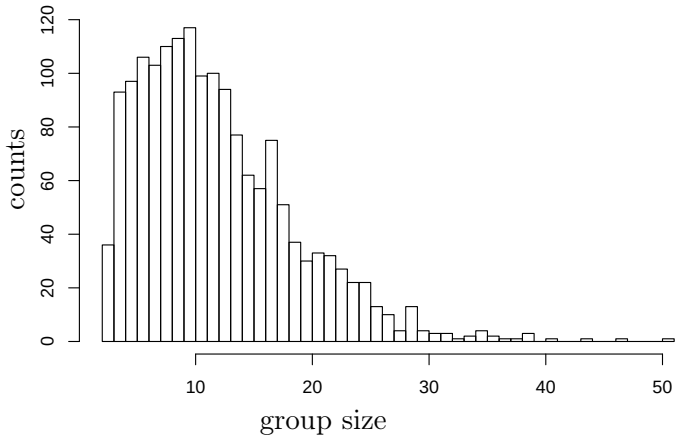


Figure 4.5 – Histogram of the size of SNP clusters based on linkage disequilibrium

Table 4.4 – p-values (in %) for the groups of SNPs passing the screening stage, either based on group-Lasso (GL) or sparse-group-Lasso (SGL). The cleaning stage relies on adaptive ridge. The group is identified by its associated gene if any. Group names in bold corresponds to regions associated to RNA and DNA HIV levels in Dalmasso et al. (2008). Starred p-values indicate selected groups by a Benjamini-Hochberg procedure at a 25% FDR level.

Associated gene	p-value		# SNPs	
	GL	SGL	GL	SGL
PSORS1C3/HCG27	8.0	**3.2***	24	12
HLA-C	11.6	**4.0***	27	17
None	13.0	–	3	0
SLC2A12	16.0	12.8	3	3
COL21A1	18.4	17.6	4	4
CYP39A1/RCAN2	20.4	–	10	0
LOC105377972	30.4	22.8	18	11
C6orf165	40.0	–	7	0
MANEA	45.2	38.8	12	7
None	59.4	54.8	2	2
MAP3K5/PEX7	79.8	71.6	12	11
LOC101928331	98.0	97.0	9	7

We applied the screen and clean procedure with group-Lasso and

sparse-group-Lasso using the groups inferred from linkage disequilibrium. The screening stage selects 12 and 9 groups with GL and SGL respectively. The groups selected by SGL belong to the subset selected by GL. Tables 4.4 and 4.5 show the results of the cleaning stage based on adaptive-ridge and ordinary least squares respectively.

When screening is based on group-Lasso, the procedure does not retain any group at the 25% FDR (with a Benjamini-Hochberg correction). When screening is based on sparse-group-Lasso, two groups remain after adaptive-ridge or ordinary least squares cleaning. These groups can be associated to gene HLA-C for one group, and to the combination of PSORS1C3 and HCG27 for the other one. Genes PSORS1C3 and HCG27 are contiguous to HLA-C on the chromosome 6. These two groups are particularly interesting because they contain 7 SNPs estimated as relevant by Dalmasso et al. (2008). Roughly half of the SNPs in these two groups is estimated irrelevant by SGL, and these eliminations are consistent with the univariate approach. SGL seems more appropriate here than GL.

OLS at cleaning yields similar results but the PSORS1C3/HCG27 group is not selected and the gene SLC2A12 GROUP replaced it. The SLC2A12 gene is not known to be associate to the viral load of HIV but it will can be interesting for future works. PSORS1C3, HCG2 and HLA-C genes are very close on the genome and they are highly correlated but their linkage disequilibrium it is not sufficient to do a unique group. However, that could explain the difficulty to the OLS to select all of relevant groups when it was too correlation as it shown in our simulation.

The present analysis fails in providing more confidence in the results. This may be explained by the fact that the number of groups remains important in comparison to the number of examples. We retain however that the results of Dalmasso et al. (2008), the ones of Chapter 3 and the present ones, derived from the same data, have few common conclusions. However, HLA-C gene and its environment have been returned by all analyses and seems thus to be a good candidate

Table 4.5 – Significance of groups based on Ordinary least squares estimates. Each groups in $\hat{\mathcal{S}}(\hat{\lambda})$ is represented by its associated gene. The p-values and the size of each groups correspond to a screening with GL or SGL. Groups in bold corresponds to regions associate to RNA and DNA HIV levels in Dalmasso et al. (2008). Groups with a star have been selected by a Benjamini-Hochberg procedure with $\alpha = 25\%$. 3 groups appearing on GL selection are not in the SGL selection.

Associated gene	Size Group	GL	Size SGL	SGL
SLC2A12	3	6.1	3	0.8*
HLA-C	27	64.8	17	**3.8***
PSORS1C3/HCG2	24	30.0	12	**12.9**
LOC105377972	18	41.1	11	28.6
MAP3K5/PEX7	12	52.8	11	34.3
MANEA	12	63.0	7	39.1
LOC101928331	9	52.5	7	44.7
COL21A1	4	74.5	4	58.0
None	2	85.0	2	82.0
C6orf165	7	34.0	-	-
CYP39A1/RCAN2	10	86.8	-	-
None	3	97.0	-	-

to explain the viral load of HIV in plasma.

Conclusion

Le travail effectué durant cette thèse a abouti à une approche permettant de réaliser, en pratique, la sélection de variables en grande dimension avec un contrôle du taux de fausses découvertes (FDR). Initialement, notre but était d'utiliser des tests statistiques "classiques" existants et de les transposer à des méthodes de régression parcimonieuses comme le Lasso et le *group-Lasso*. Nous avons fait nos choix méthodologiques de manière à ce qu'ils s'appliquent de manière générale, c'est à dire sans contraintes sur les dimensions, sur une famille de pénalités assez large. Pour la même raison, une grande attention à été portée sur l'efficacité algorithmique.

Nous avons vu dans le cadre de la régression ridge qu'il est difficile de déterminer les distributions que suivent les statistiques construites des estimateurs pénalisés. Si la forme de ces distributions était connue, il resterait à déterminer leurs paramètres. Des solutions en petite dimension existent mais elles ne sont pas applicables en grande dimension (Halawa and El Bassiouni 1999). D'autres existent pour circonvenir, indépendamment de la notion de dimension, le problème du biais (Bühlmann 2013) mais pas celui de la colinéarité qui rend injustifié l'usage de statistiques de Student ou de Fisher.

Partant de constat, nous avons proposé une approche basée sur les permutations et la statistique de Fisher. Cette approche permet de simuler la distribution des statistiques sous l'hypothèse nulle, et d'obtenir ainsi un test approximatif. Dans le cadre de nos simulations, ce test se comporte bien, que ce soit en grande dimension ou sur les pénalités fortes. Contrairement à d'autres méthodes de rééchantillonnage que nous avons testées, les biais sur les estimateurs et sur les

résidus n'entraînent pas d'erreur. Ceci est du à la comparaison de deux modèles où la différence de ces biais apparaît être suffisamment faible. Par ailleurs, bien que les approches de rééchantillonage aient par nature un fort coût calculatoire, nous avons pu le limiter fortement par l'utilisation de décompositions classiques d'inverses de matrices.

Nous avons pu définir une liste de pénalités spécifiques permettant d'implémenter le Lasso et ses variantes, par le biais de la régression ridge, dans la lignée des travaux de Grandvalet (1998). Le principe de la l'*adaptive ridge*, avec laquelle nous pouvons faire un parallèle avec l'*adaptive Lasso*, est particulièrement bien adapté aux procédures en deux étapes en utilisant ces pénalités. Dans notre cas, nous nous sommes basés sur le protocole "screen and clean" de Wasserman and Roeder (2009), dont la première étape de criblage de variables par le Lasso est suivie d'une seconde étape de nettoyage par un test de Student basé sur l'estimateur des moindres carrés.

Durant ces travaux nous avons étendu cette procédure aux variantes du Lasso (*group-Lasso*, ...) pour la pré-sélection. Cette extension est rendue possible par l'utilisation de l'*adaptive ridge* dans l'étape de nettoyage. Les tests de cette étape utilisent notre approche de permutation développée sur la régression ridge. La procédure d'origine faisait un usage beaucoup moins intensif des données de l'étape de pré-sélection, en ne gardant de cette étape que le support estimé, ce qui nous semblait dommageable. Le fait d'utiliser l'*adaptive ridge* en lieu et place des moindres carrés permet d'avoir un transfert plus complet de l'information (poids des variables, structure des groupes) entre les deux étapes. Nos simulations et une analyse du cas orthogonal ont montré que l'utilisation de l'*adaptive ridge* offrait toujours une meilleure sensibilité que les moindres carrés pour un même contrôle de l'erreur.

Un point essentiel sur la définition des variables explicatives, est apparu en filigrane tout au long cette thèse. En effet, cette définition dépend d'une part de la question biologique associée (recherche de cibles précises, de régions d'intérêt, ...) et d'autre part de notre

capacité à différencier les corrélations directes de celles indirectes. En pratique, nous ne sommes pas assurés d'observer les variables causales. De plus, dans un procédé en deux étapes, nous ne sommes pas sûrs qu'elles soient sélectionnées durant le criblage (*screening*). La formalisation du problème basée sur la couverture de Markov nous semble définir le problème de manière satisfaisante. Par ailleurs, les approches s'intéressant directement aux groupes de variables semblent être une voie intéressante car en élargissant la notion de variable d'intérêt on perdra en précision mais on gagnera en confiance. Ceci permet aussi de circonvenir le problème du masquage des variables corrélées qui participent grandement à l'instabilité du Lasso. Toutefois se pose la question de la définition des groupes qui n'est pas triviale.

Les travaux effectués dans le cadre de cette thèse et plus particulièrement la partie sur l'*adaptive ridge* (chapitre 3) ont été diffusés sous la forme d'un article dans la revue internationale *Statistics and Computing* (Becu et al. 2015b) et de plusieurs communications dans des congrès nationaux comme les journées Modélisation Aléatoire et Statistique (MAS'14), internationaux comme *Statistical Methods for Post Genomic Data* (SMPGD'15) et dans les actes de la conférence*Computational Intelligence in Bioinformatics and Computational Biology* (CICBC'15) (Becu et al. 2015a).

Les objectifs de cette thèse ont été en grande partie atteints même si nous regrettons ne pas avoir réussi à prouver, autrement que par des simulations, la pertinence de nos propositions. Toutefois, ces propositions nous semblent avoir un intérêt pratique certain. Les travaux à venir ont, entre autres, vocation à consolider les bases théoriques. Il serait également intéressant de tester et d'adapter notre méthode à l'inférence de graphes. Il est à noter qu'un package R, nommé *ridgeAdap*, est actuellement accessible sur ma page personnelle (https://www.hds.utc.fr/~becujean/dokuwiki/fr/accueil) et sera mis à terme, en version finale sur les serveurs du CRAN.

Bibliography

M. J. Anderson and P. Legendre. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, 62(3): 271–303, 1999.

M. J. Anderson and J. Robinson. Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88, 2001.

F. R. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM, 2008.

R. F. Barber and E. Candes. Controlling the false discovery rate via knockoffs. *arXiv preprint arXiv:1404.5609*, 2014.

J.-M. Becu, C. Ambroise, Y. Grandvalet, and C. Dalmasso. Significance testing for variable selection in high-dimension. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, pages 1–8, Aug 2015a. doi: 10.1109/CIBCB.2015.7300313.

J.-M. Becu, Y. Grandvalet, C. Ambroise, and C. Dalmasso. Beyond support in two-stage variable selection. *Statistics and Computing*, pages 1–11, 2015b. ISSN 0960-3174. doi: 10.1007/s11222-015-9614-1. URL http://dx.doi.org/10.1007/s11222-015-9614-1.

A. Beinrucker, Ü. Dogan, and G. Blanchard. Extensions of stability selection using subsamples of observations and covariates. *arXiv preprint arXiv:1407.4916*, 2014.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

Y. Benjamini and Y. Hochberg. 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, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

Y. Benjamini and D. Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA., 1984.

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19:1212–1242, 2013.

P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, 2013.

A. Chatterjee and S. Lahiri. Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12):4497–4509, 2010.

A. Chatterjee and S. N. Lahiri. Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259, 2013.

J. Chiquet, Y. Grandvalet, and C. Charbonnier. Sparsity with sign-coherent groups of variables via the cooperative-lasso. *The Annals of Applied Statistics*, 6(2):795–830, 2012.

D. R. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

A. Crivelli, L. Firinguetti, R. Montaño, and M. Muñóz. Confidence intervals in ridge regression by bootstrapping the dependent variable: a simulation study. *Communications in Statistics-Simulation and Computation*, 24(3):631–652, 1995.

E. Cule, P. Vineis, and M. De Lorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(372):1–15, 2011.

C. Dalmasso, P. Broët, and T. Moreau. A simple procedure for estimating the false discovery rate. *Bioinformatics*, 21(5):660–668, 2005.

C. Dalmasso, W. Carpentier, L. Meyer, C. Rouzioux, C. Goujard, M.-L. Chaix, O. Lambotte, V. Avettand-Fenoel, S. Le Clerc, L. Denis de Senneville, C. Deveau, F. Boufassa, P. Debre, J.-F. Delfraissy, P. Broet, and I. Theodorou. Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS genome wide association 01 study. *PLoS One*, 3(12): e3907, 2008.

A. Dehman, C. Ambroise, and P. Neuvial. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC bioinformatics*, 16(1):148, 2015.

L. DO Q. Numerically efficient methods for solving least squares problems. 2012.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2015.

D. Freedman and D. Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4): 292–298, 1983.

D. A. Freedman. A note on screening regression equations. *The American Statistician*, 37(2):152–155, 1983.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. Technical report, arXiv, 2010.

Y. Ge, S. C. Sealfon, and T. P. Speed. Some step-down procedures controlling the false discovery rate under dependence. *Statistica Sinica*, 18(3):881, 2008.

Y. Grandvalet. Least absolute shrinkage is equivalent to quadratic penalization. In L. Niklasson, M. Bodén, and T. Ziemske, editors, *ICANN'98*, volume 1 of *Perspectives in Neural Computing*, pages 201–206. Springer, 1998.

Y. Grandvalet and S. Canu. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11 (NIPS 1998)*, pages 445–451. MIT Press, 1999.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

A. M. Halawa and M. Y. El Bassiouni. Tests of regressions coefficients under ridge regression models. *Journal of Statistical Computation and Simulation*, 65(1):341–356, 1999.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1990.

S. Hochreiter, D.-A. Clevert, and K. Obermayer. A new summarization method for affymetrix probe level data. *Bioinformatics*, 22(8):943–

949, 2006. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/22/8/943.

A. E. Hoerl and R. W. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 1970.

J. Huang, J. L. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, 36(2):587–613, 2008.

V. A. Huynh-Thu, Y. Saeys, L. Wehenkel, and P. Geurts. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics (Oxford, England)*, 28(13): 1766, 2012.

K. Jong, J. Mary, A. Cornuéjols, E. Marchiori, and M. Sebag. Ensemble feature ranking. In *Knowledge Discovery in Databases: PKDD 2004*, pages 267–278. Springer, 2004.

F. E. Kennedy. Randomization tests in econometrics. *Journal of Business & Economic Statistics*, 13(1):85–94, 1995.

Y.-T. Lin and W.-C. Lee. Importance of presenting the variability of the false discovery rate control. *BMC Genetics*, 16(1):97, 2015. ISSN 1471-2156. doi: 10.1186/s12863-015-0259-z. URL http://www.biomedcentral.com/1471-2156/16/97.

J. Listgarten and D. Heckerman. Determining the number of non-spurious arcs in a learned DAG model. *Proeedings of UAI*, 2007.

H. Liu and B. Yu. Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics*, 7:3124–3169, 2013.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

M. E. Lopes. A residual bootstrap for high-dimensional regression with near low-rank designs. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3239–3247. Curran Associates, Inc., 2014.

B. F. J. Manly. *Randomization, bootstrap and Monte Carlo methods in biology*, volume 70. CRC Press, 2006.

N. Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374 – 393, 2007.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 417–473, 2010.

N. Meinshausen, L. Meier, and P. Bühlmann. p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

R. L. Obenchein. Classical F-tests and confidence regions for ridge regression. *Technometrics*, 19(4):429–439, 1977.

A. Rakotomamonjy. Variable selection using SVM-based criteria. *The Journal of Machine Learning Research*, 3:1357–1370, 2003.

R. K. Rayner. Bootstrap inversion of edgeworth expansions for non-parametric confidence intervals. *Statistics & Probability Letters*, 8 (3):201–206, 1989.

S. Sartori. *Penalized Regression: bootstrap confidence intervals and variable selection for high dimensional data sets.* PhD thesis, Université de Milan, 2010.

L. C. Snouber, S. Jacques, M. Monge, C. Legallais, and E. Leclerc. Transcriptomic analysis of the effect of ifosfamide on MDCK cells cultivated in microfluidic biochips. *Genomics*, 100(1):27–34, 2012.

J. D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of statistics*, pages 2013–2035, 2003.

C. J. Ter Braak. Permutation versus bootstrap significance tests in multiple regression and anova. In *Bootstrapping and related techniques*, pages 79–85. Springer Berlin Heidelberg, 1992.

R. J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996.

L. Wasserman and K. Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009.

H. Yi, P. Breheny, N. Imam, Y. Liu, and I. Hoeschele. Penalized multimarker vs. single-marker regression methods for genome-wide association studies of quantitative traits. *Genetics*, 199(1):205–222, 2015.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 67(2):301–320, 2005.