THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Institut National des Sciences Appliquées de Toulouse (INSA Toulouse)

Discipline ou spécialité :
Mathématiques appliquées

Présentée et soutenue par
Florian ROHART
le : 7 décembre 2012

Titre :
Prédiction phénotypique et sélection de variables en grande dimension dans les modèles linéaires et linéaires mixtes

---

Jury

David CAUSEUR
Marie-Luce TAUPIN
Jean-Michel LOUBES
Béatrice LAURENT-BONNEAU
Christian LAVERGNE
Florence PHOCAS
Philippe BESSE
Magali SAN CRISTOBAL

École doctorale :
Mathématiques Informatique Télécommunications (MITT)

Unité de recherche :
UMR 0444 et UMR 5219

Directeur(s) de thèse :
Béatrice LAURENT-BONNEAU et Magali SAN CRISTOBAL

Rapporteurs :
David CAUSEUR et Christian LAVERGNE

Prédiction phénotypique et sélection de variables en grande dimension dans les modèles linéaires et linéaires mixtes

—

Phenotypic prediction and variable selection in high dimensional linear and linear mixed models

Florian Rohart

2012

"Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid."
Albert Einstein

"It's not who you are underneath, it's what you do that defines you."
Batman Begins

Remerciements

Quand on m'a demandé "tu comptes mettre quoi dans tes remerciements ? Tu vas remercier qui ? Tu vas me remercier ?", je me suis rendu compte qu'il fallait réfléchir à la question, le problème étant de savoir qui remercier et comment tourner les choses pour n'oublier personne (oui je connais des gens très susceptibles, ils se reconnaîtront :).

Au fil des réflexions, on se rend compte que l'on en arrive à écrire les remerciements d'une thèse après 8 années d'études supérieures (ou 9 si on tient un carnet de compte) grâce à beaucoup de personnes ! Si on commence par le tout début, il faut dire un grand merci à mes années et mes années de prépa, je ne pense vraiment pas que j'en serais arrivé là sans vous ! Vous m'avez donné goût aux mathématiques, les vrais, avec des epsilons (positifs), vous m'avez apporté la rigueur et la persévérance, et vous m'avez surtout apporté des amis dignes de ce nom ! Donc un grand merci à vous et surtout aux très bons profs que j'ai eu durant cette période qui restera sûrement une des périodes les plus jouissives de ma vie (contrairement à ce qu'on pourrait penser quand on entend les commentaires de certains élèves sur la prépa...). Je me souviens encore de ces pizzerias que l'on faisait tous ensembles, élèves et profs, dans une superbe ambiance, où petite anecdote qui pourrait être utile à quelques uns : les profs passaient toujours en dernier et il ne restait plus jamais rien à payer ;). On leur devait bien ça en récompense de tout ce qu'ils nous apportaient ! Et l'internat, parlons-en, ça allait des devoirs maisons que l'on faisait le mercredi après-midi regroupé sur un bureau aux soirées PES (et pas que les soirées, 10min de pause entre deux cours ? pile le temps d'un match :). C'est grâce à tous ces moments inoubliables qu'une grande majorité d'entre nous se retrouve à Toulouse au moment où j'écris ces quelques lignes (et sûrement un peu grâce à l'attractivité du sud-ouest pour les gens d'en haut que nous étions :).

Je n'oublie bien sûr pas les amis de la fac ! Même si la majorité s'est destiné à la profession d'enseignant (on ne vous en veut pas), quelques uns ont poursuivi sur la voie de la recherche ! Ce qui donna de bonnes petites soirées entre "faqueux".

Même si la vie n'est pas toujours rose et remplie de bons côtés, l'entourage personnel mais aussi professionnel a sa grande importance dans le moral quotidien (*Prenez la vie du bon côté, Riez, sautez, dansez, chantez*). Je tiens donc à remercier mon staff technique (donc en gros mes deux directrices de thèses), je n'en serais jamais arrivé là sans lui (elles) : premièrement parce que je n'aurais jamais commencé cette thèse ; deuxièmement parce que l'encadrement était parfait (bon, je reconnais que je n'en ai pas connu d'autre donc que tout ça est assez subjectif :)) avec juste ce qu'il faut de "c'est pas terrible", "c'est un peu mieux mais c'est pas encore ça", et "on va dire que ça va". Elles m'ont soutenu, encouragé et guidé pendant ces trois ans, toujours avec le sourire. Mesdames, c'était un plaisir de travailler et d'apprendre avec vous !

Il ne faut pas oublier les membres de mon comité de thèse qui m'ont soutenu et aiguillé sur les bonnes voies à l'occasion de nos rencontres ; les rapporteurs de cette thèse pour leurs avis pertinents sur mes travaux, les membres du jury qui, je l'espère, sauront voir en moi

le futur docteur (haaa *Sauron*, pardon...).

Je n'oserai oublier les collègues du foot de l'adas-INRA, merci pour tous ces matchs qui sont de vrais moments de détente dont on a tous besoin ! Et merci de m'avoir laissé la place de meilleur buteur. . . ;).
Et enfin ma famille qui même en étant loin voire très loin ne m'a pas oublié (le nooooord en langue "toulousaine" -au dessus de Bordeaux donc), les collègues de travail que j'ai côtoyés pendant ces trois années avec plus particulièrement mes co-bureaux successifs sans qui la thèse n'aurait pas eu le même goût j'en suis sûr.
Et pour finir je te remercie toi, lecteur, et j'espère que tu prendras autant de plaisir à lire ce travail que j'en ai eu à l'accomplir.

En espérant vous revoir le 22 décembre 2012 (pour les personnes ne comprenant pas : la veille est sensé marquée la fin du ou d'un monde, sauf à Bugarach...), parce que finir Docteur c'est quand même pas mal,

Florian

Résumé

Les nouvelles technologies permettent l'acquisition de données génomiques et post-génomiques de grande dimension, c'est-à-dire des données pour lesquelles il y a toujours un plus grand nombre de variables mesurées que d'individus sur lesquels on les mesure. Ces données nécessitent généralement des hypothèses supplémentaires afin de pouvoir être analysées, comme une hypothèse de parcimonie pour laquelle peu de variables sont supposées influentes. C'est dans ce contexte de grande dimension que nous avons travaillé sur des données réelles issues de l'espèce porcine et de la technologie haut-débit, plus particulièrement le métabolome obtenu à partir de la spectrométrie RMN et des phénotypes mesurés post-mortem pour la plupart. L'objectif est double : d'une part la prédiction de phénotypes d'intérêt pour la production porcine et d'autre part l'explicitation de relations biologiques entre ces phénotypes et le métabolome. On montre, grâce à une analyse dans le modèle linéaire effectuée avec la méthode Lasso, que le métabolome a un pouvoir prédictif non négligeable pour certains phénotypes importants pour la production porcine comme le taux de muscle et la consommation moyenne journalière. Le deuxième objectif est traité grâce au domaine statistique de la sélection de variables. Les méthodes classiques telles que la méthode Lasso et la procédure FDR sont investiguées et de nouvelles méthodes plus performantes sont développées : nous proposons une méthode de sélection de variables en modèle linéaire basée sur des tests d'hypothèses multiples. Cette méthode possède des résultats non asymptotiques de puissance sous certaines conditions sur le signal. De part les données annexes disponibles sur les animaux telles que les lots dans lesquels ils ont évolués ou les relations de parentés qu'ils possèdent, les modèles mixtes sont considérés. Un nouvel algorithme de sélection d'effets fixes est développé et il s'avère beaucoup plus rapide que les algorithmes existants qui ont le même objectif. Grâce à sa décomposition en étapes distinctes, l'algorithme peut être combiné à toutes les méthodes de sélection de variables développées pour le modèle linéaire classique. Toutefois, les résultats de convergence dépendent de la méthode utilisée. On montre que la combinaison de cet algorithme avec la méthode de tests multiples donne de très bons résultats empiriques. Toutes ces méthodes sont appliquées au jeu de données réelles et des relations biologiques sont mises en évidence.

Abstract

Recent technologies have provided scientists with genomics and post-genomics high-dimensional data ; there are always more variables that are measured than the number of individuals. These high dimensional datasets usually need additional assumptions in order to be analyzed, such as a sparsity condition which means that only a small subset of the variables are supposed to be relevant. In this high-dimensional context we worked on a real dataset which comes from the pig species and high-throughput biotechnologies. Metabolomic data has been measured with NMR spectroscopy and phenotypic data has been mainly obtained post-mortem. There are two objectives. On one hand, we aim at obtaining good prediction for the production phenotypes and on the other hand we want to pinpoint metabolomic data that explain the phenotype under study. Thanks to the Lasso method applied in a linear model, we show that metabolomic data has a real prediction power for some important phenotypes for livestock production, such as a lean meat percentage and the daily food consumption. The second objective is a problem of variable selection. Classic statistical tools such as the Lasso method or the FDR procedure are investigated and new powerful methods are developed. We propose a variable selection method based on multiple hypotheses testing. This procedure is designed to perform in linear models and non asymptotic results are given under a condition on the signal. Since supplemental data are available on the real dataset such as the batch or the family relationships between the animals, linear mixed models are considered. A new algorithm for fixed effects selection is developed, and this algorithm turned out to be faster than the usual ones. Thanks to its structure, it can be combined with any variable selection methods built for linear models. However, the convergence property of this algorithm depends on the method that is used. The multiple hypotheses testing procedure shows good empirical results. All the mentioned methods are applied to the real data and biological relationships are emphasized.

Table des matières

1 Introduction

Les nouvelles technologies permettent l'acquisition de données de plus en plus complexes et de plus en plus gigantesques par leur taille. Ce phénomène est visible dans la plupart des secteurs scientifiques, comme l'aéronautique, l'espace, la médecine ou encore la biologie. De nos jours, l'acquisition de données est beaucoup plus facile et rapide que leurs analyses poussées. Des problèmes évidents découlent de cette "course à l'armement" tels que la sauvegarde de toutes ces données et bien sûr le besoin en moyens humains et donc en statisticiens pour les analyser en profondeur. Prenons en exemple le cas de la biologie. Depuis l'avènement des technologies haut-débit, des données complexes et précises sont récoltées par les scientifiques. Les progrès scientifiques et technologiques fournissent maintenant la capacité de séquencer un génome complet, mais aussi d'obtenir des données post-génomiques comme la mesure d'une grande partie du transcriptome à l'aide de puce ou de séquenceurs (RNA-seq) ou le métabolome à l'aide de la spectrométrie de résonance magnétique. Ces données sont généralement de très grande dimension, quelques dizaines de milliers de variables mesurées pour les données transcriptomiques et jusqu'à plusieurs milliards pour les données génomiques. Ceci pose le problème de la très grande dimension puisqu'il y aura toujours moins d'observations que de paramètres. Les outils d'analyse traditionnels, comme la régression linéaire, nécessitent généralement plus d'observations que de variables, ce qui signifie que pour analyser des données transcriptomiques dans ce cadre, il faudrait des dizaines de milliers d'observations, et donc des dizaines de milliers d'individus. Des outils moins traditionnels sont donc nécessaires, permettant l'analyse de données complexes dans le cas où il y a plus voire beaucoup plus de variables que d'individus. Des hypothèses sont généralement faites sur la réalité sous-jacente du modèle afin de parer aux problèmes de grande dimension, comme par exemple l'hypothèse de parcimonie selon laquelle seul un nombre limité de variables est important. Il y a toutefois des cas où le rapport entre le nombre de variables et le nombre d'observations est tel qu'il est impossible d'identifier ces quelques variables pertinentes.

Les données génomiques et post-génomiques issues de technologies haut-débit permettent d'avoir un regard sur la cascade physiologique qui va du gène au phénotype d'un individu. Des questions naturelles émergent de cette masse de données telles que "Comment peut-on prédire des phénotypes d'intérêt économique et les gérer sur la base de la connaissance de leur information moléculaire ?" ou encore "Comment expliciter les relations entre des phénotypes et des données génomiques ou post-génomiques ?".

L'inscription de ce travail de recherche dans le projet ANR DéLiSus permet l'accès à un jeu de données réelles provenant de l'espèce porcine, espèce d'importance économique cruciale et première source mondiale de protéine dans l'alimentation humaine. Le projet DéLiSus est un projet intégré ayant pour but l'étude de la variabilité haplotypique du génome porcin à haute densité. L'analyse des haplotypes permet une analyse très détaillée de la diversité génétique des races porcines et la détection de traces de sélection révélant des régions génomiques ayant répondu à la sélection. Des données génomiques et post-

génomiques haut débit ont été récoltées sur plusieurs centaines d'animaux. Les objectifs du projet DéLiSus avaient trait à la caractérisation fine des principales races porcines françaises, au niveau génétique et au niveau phénotypique. Cette thèse s'est focalisée sur les phénotypes "fins", et en particulier le métabolome, en lien avec des phénotypes "finaux" de production.

1.1 Objectifs

Ce travail a été motivé par des applications agronomiques et a nécessité le développement de nouvelles méthodes statistiques. C'est donc toujours dans cet esprit que les travaux de cette thèse ont été envisagés : le travail théorique fourni devant servir un but précis et répondre à une question appliquée. Dans ce cadre, nous avons soulevé deux questions principales :

- Comment peut-on prédire des phénotypes d'intérêt économique pour la filière porcine et les gérer sur la base de la connaissance de leur information moléculaire ?
- Comment expliciter les relations entre ces phénotypes et des données génomiques ou post-génomiques ?

Pour répondre à ces questions, les premières données que nous avons eues à notre disposition ont été des données métabolomiques et phénotypiques de plusieurs centaines d'animaux. La grande majorité de ce manuscrit se focalisera sur ces données, sauf mention contraire. L'objectif est ici double.

Le premier objectif est un objectif de prédiction. Nous cherchons à prédire au mieux un phénotype donné, c'est-à-dire réussir à être le plus proche possible de la valeur du phénotype en ayant seulement accès à des données métabolomiques. Cet objectif peut avoir des implications importantes dans le monde de l'élevage en fonction de la qualité de la prédiction des phénotypes. En effet, des phénotypes tels que des indicateurs de qualité de la viande, le poids de jambon ou le taux de muscle, qui sont des phénotypes qui influencent directement le paiement des éleveurs, peuvent aider la filière porcine s'ils sont bien prédits, à partir par exemple d'une simple prise de sang donnant accès à des données métabolomiques. On peut imaginer que le suivi d'un animal le long de sa vie soit simplement fait à partir de prises de sang régulières qui détermineront le moment opportun où l'éleveur devra se séparer de cet animal.

Le second objectif consiste à expliciter des mécanismes biologiques sous-jacents à un phénomène donné. Par exemple, pourquoi cet individu est-il beaucoup plus gras que cet autre individu soumis aux mêmes conditions d'élevage ? Qu'est-ce qui dans leurs génomes, dans leurs métabolismes, permet de donner une réponse ? Sur les données réelles, cela revient à avoir comme objectif d'identifier quelles variables parmi l'ensemble des variables métabolomiques expliquent le mieux le phénotype considéré, c'est-à-dire à mettre en évidence des relations potentielles entre le métabolome et un phénotype. Cette question biologique s'apparente en statistique à la question de la sélection de variables. La sélection de variables est la question majeure de ce manuscrit : répondre à ce problème permet de comprendre des

mécanismes biologiques mis en jeu. Contrairement au problème de prédiction dans lequel on s'autorise à prédire à l'aide de toutes les variables ainsi qu'à opérer des transformations sur ces variables, la sélection s'effectue généralement sur les données brutes et vise à obtenir un faible nombre de variables afin de faciliter l'interprétation biologique et ainsi de pouvoir répondre à notre second objectif. La sélection de variables est un sujet d'actualité qui a de multiples applications directes, notamment en biologie avec par exemple la recherche de bio-marqueurs ou encore la sélection génomique (qui consiste à prédire la valeur génétique d'animaux dès leur naissance, sans attendre la collecte de phénotypes).

Détaillons les données dont nous disposons avant de faire un inventaire non exhaustif des méthodes existantes qui répondent à nos objectifs.

1.2 Les données

Les animaux du projet ont été suivis dans la station de contrôle française basée au Rheu, en 2007 et 2008 dans huit bandes différentes. On entend par bande le fait que les animaux sont élevés en lots et qu'ils sont soumis aux mêmes conditions intra-groupe (condition climatiques, nourriture, ...). Les individus étaient regroupés par 12 du début de la période de contrôle -à environ un âge de 10 semaines- jusqu'au jour précédant leur mort -à environ 110 kg, soit en moyenne à 172 jours-. La plupart de ces individus étaient apparentés, demi-frères pour la majorité d'entre eux (même père mais mère différente). Le génome, le transcriptome, le métabolome et certains phénotypes ont été recueillis sur ces animaux. Il est à noter qu'à cause des coûts de production des différentes données, tous les types de données n'ont pas été recueillis sur tous les animaux.

1.2.1 Le métabolome

Le métabolome est constitué de l'ensemble des métabolites -petites molécules telles que le glucose ou la créatinine- contenus dans un système biologique donné (cellules ou fluides biologiques tels que les urines ou le plasma). Les données métabolomiques étudiées ici ont été obtenues sur des échantillons de plasma prélevés en moyenne à un poids de 60 kg, grâce à la technologie de la spectroscopie de résonance magnétique nucléaire (spectroscopie RMN). Cette technique repose sur le fait que les molécules n'ont pas toutes la même fréquence de résonance. Les données issues de la spectroscopie RMN se présentent sous forme de spectre. Un travail préalable est fait sur ces spectres afin d'obtenir des données "propres" : les pics sont alignés et la ligne de base est corrigée, puis les spectres sont discrétisés en "buckets" qui sont alors normalisés par rapport à l'intensité totale du signal de chaque spectre. Ce nettoyage technique des données est expliqué plus en détail dans la Section 2.2. Un exemple de spectre est présenté en Figure 1.
Pour les interprétations biologiques, il est important de noter que certains 'buckets' (points du spectre discrétisé) signent la présence d'un ou de plusieurs métabolites (connus ou pas) et que certains métabolites peuvent "résonner" sous plusieurs buckets. Les données

finales comportent 375 variables (buckets) pour 658 individus de 8 races différentes : Duroc, Duoschan, Musclor, Tai Zumu, Large White type femelle, Large White type mâle, Landrace et Piétrain. Nous nous sommes focalisés sur trois grandes races, Large White type femelle, Landrace et Piétrain, car elles ont fait l'objet d'un échantillonnage de plus grande taille (au minimum 121 individus) pour un total de 506 individus.

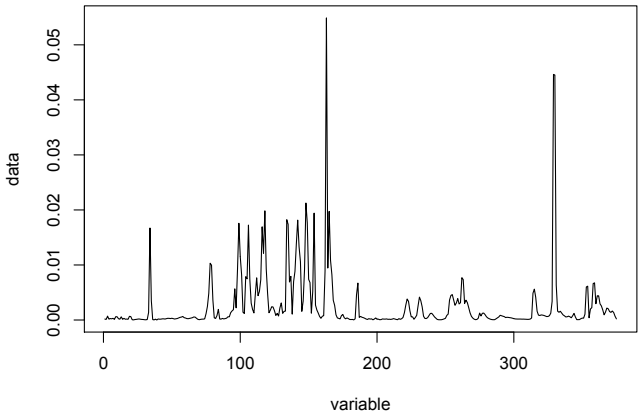


FIGURE 1 – Données métabolomiques d'un individu.

Il est important de noter que le métabolome est dépendant de l'environnement : les taux de métabolites sont modifiés en fonction de l'état physiologique, développemental, ou pathologique d'une cellule, d'un tissu, d'un organe ou d'un organisme. Le métabolome est donc une variable dynamique, vouée à changer dans le temps. Il pourrait s'apparenter à une photo de l'individu à un instant précis.

1.2.2 Le phénome

Le phénome se compose de 27 phénotypes recueillis post-mortem pour la plupart, parmi lesquels des taux de muscle, l'épaisseur de gras à différents endroits, des indicateurs de qualité de viande, etc. Ces phénotypes se regroupent en 5 groupes : poids de l'animal, croissance, poids de carcasse, composition de la carcasse et indicateurs de qualité de la viande. Les phénotypes sont détaillés dans la Table 1.

	phénotype	signification
Poids	LWETP	Poids de l'animal en fin de période de contrôle ; kg
	LWS	Poids vif de l'animal (à jeun) ; kg
Croissance	ADG	Gain Moyen Quotidien ; g/j
	FCR	Indice de Consommation ; kg/kg
	DFI	Consommation Moyenne Journalière de l'animal = FCR*ADG ; kg/j
Poids de carcasse	CW	Poids net avec tête ; kg
	CWwtH	Poids net sans tête ; kg
	HCW	Poids de la demi carcasse droite ; kg
	DP	Rendement de carcasse
Composition de la carcasse	hamW	Poids du jambon de la demi carcasse droite ; kg
	loinW	Poids de poitrine de la demi carcasse droite ; kg
	bfW	Poids d'épaule de la demi carcasse droite ; kg
	shW	Poids de la longe de la demi carcasse droite ; kg
	beW	Poids de la bardière de la demi carcasse droite ; kg
	LMP	Taux de Muscle des Pièces estimé à l'aide des poids de jambon, de longe et de bardière dans la demi carcasse
	Com.LMP	Taux de Muscle des Pièces commercial (autre estimation que LMP)
	Length	Longueur de la carcasse ; mm
	BFsh	Epaisseur de gras à la fente au niveau des reins ; mm
	BFlr	Epaisseur de gras à la fente au niveau du dos ; mm
	BFhj	Epaisseur de gras à la fente au niveau du cou ; mm
	mBF	Moyenne des 3 épaisseurs de gras à la fente = (BFsh+BFlr+BFhj)/3
Qualité de la viande	pH24	pH ultime (24h post mortem) du muscle demi membraneux
	L*	Mesure L* du muscle fessier superficiel (minolta)
	a*	Mesure a du muscle fessier superficiel (minolta)
	b*	Mesure b du muscle fessier superficiel (minolta)
	WHC	Temps d'imbibition d'un morceau de papier pH sur le muscle fessier superficiel (=indicateur de la capacité de rétention d'eau du muscle) ; en dizaines de secondes
	MQI	Rendement technologique estimé (=indicateur de la qualité technologique du jambon)

TABLE 1 – **Liste et signification des 27 phénotypes**

1.3 Prédiction et sélection de variables en grande dimension

Mettre à jour des relations entre des variables explicatives (gènes, métabolites, etc.) et une observation (phénotype ou autres) est un problème majeur en biologie, notamment pour la recherche de marqueurs biologiques. Avec l'apparition des données de grande di-

mension, on cherche souvent à déterminer un petit ensemble de variables qui expliquent l'observation quasiment aussi bien que toutes les variables mais qui permet de mieux prédire l'observation. En effet, si toutes les variables sont considérées comme étant pertinentes, on risque de se positionner dans un contexte de sur-apprentissage. Un point commun de toutes les données sur lesquelles nous avons travaillé est la grande dimension : le nombre p de variables excède le nombre n d'individus. Nous allons nous consacrer dans ce paragraphe au modèle linéaire.

La sélection de variables peut-être interprétée comme une ramification de la sélection de modèle. En effet sélectionner les bons paramètres parmi une collection de p paramètres revient à sélectionner le bon modèle parmi une collection de 2^p modèles. De nombreux travaux de recherche dans le domaine de la sélection de modèles ont été développés ces dernières années, en particulier dans le cadre de modèles gaussiens. Birgé and Massart (2001) ont proposé d'opérer la sélection de modèles à partir d'un critère pénalisé, mais les auteurs travaillent à variance connue, ce qui est rarement le cas en pratique. Baraud et al. (2009) ont alors considéré la sélection de modèle gaussien à variance inconnue en proposant un critère de choix de modèles pénalisé.

La sélection de variables en elle-même a connu un regain d'activité à la fin des années 1990 avec l'apparition de la méthode Lasso par Tibshirani (1996). Le Lasso est une méthode basée sur un critère pénalisé très simple qui permet de faire de la sélection de variables dans un modèle linéaire, cette méthode est applicable lorsqu'il y a plus de variables explicatives que d'observations. Le Lasso a reçu beaucoup d'attention et de nombreux résultats théoriques sont disponibles, comme la consistance (Zhao and Yu, 2006), des résultats sur la sélection de variables dans des graphes gaussiens (Meinshausen and Bühlmann, 2006) ou des résultats de consistance lorsque la méthode est combiné à un test de Student dans la procédure "screan and clean" (Wasserman and Roeder, 2009). Le Lasso possède également de nombreuses extensions comme le Bolasso (Bach, 2009), l'adaptive Lasso (Zou, 2006; Huang et al., 2008) ou le group Lasso (Yuan and Lin, 2007; Chesneau and Hebiri, 2008). Les résultats pratiques sur ces différentes méthodes étant relatifs, Meinshausen and Bühlmann (2010) ont introduit de la stabilité par un 'randomized Lasso'. La méthode Lasso fonctionne en grande dimension, mais elle n'a pas été construite à cette fin contrairement au Dantzig selector (Candes and Tao, 2007). Néanmoins, Bickel et al. (2009) montrent que le Lasso et le Dantzig selector se comportent de la même façon sous une condition de parcimonie.

Toutes ces méthodes sont basées sur une pénalisation ℓ^1, mais la combinaison d'une pénalité ℓ^1 avec une pénalité ℓ^2 a été envisagée par Zou and Hastie (2005) sous le nom d'elastic net, la pénalité ℓ^2 étant connue sous le nom de régularisation Tikhonov (ou régression ridge lorsqu'elle est appliquée en régression).

Le critère pénalisé n'est pas le seul moyen de faire de la sélection de variables. En effet, les tests multiples peuvent aussi s'avérer utiles et ils sont notamment employés à travers la procédure FDR (Bunea et al., 2006) qui a été développée par Benjamini and Hochberg (1995). Cette procédure est largement utilisée en pratique pour découvrir des

gènes différentiels, entre deux conditions par exemple. Dans le cas de données de grande dimension fortement corrélées, comme c'est le cas avec des données de puces, le package R Factor Analysis for Multiple Testing (FAMT) est tout adapté (Causeur et al., 2011).

Introduisons le modèle linéaire avant d'expliciter quelques méthodes de sélection de variables dans ce modèle.

1.3.1 Le modèle linéaire

Nous considérons le modèle linéaire suivant :

$$Y = X\beta + \epsilon, \tag{1}$$

où Y est un vecteur de données observées de longueur n (un phénotype par exemple), $X = (X_1, \ldots, X_p)$ est la matrice de taille $n \times p$ des p variables mesurées sur les n individus (comme les données métabolomiques). Pour tout i, X_i est le vecteur de $\mathbb{R}^n$ associé à la $i^{ème}$ variable. $\beta = (\beta_1, \ldots, \beta_p)$ est le vecteur des coefficients et ϵ est un bruit gaussien : $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ où σ est un paramètre positif inconnu et I_n est la matrice identité de $\mathbb{R}^n$.

L'estimateur classique d'un modèle linéaire est l'estimateur des moindres carrés ('Ordinary Least Squares') :

$$\beta_{OLS} = \underset{\beta \in \mathbb{R}^p}{Argmin} \left\{ ||Y - X\beta||_2^2 \right\}, \tag{2}$$

où $||.||_2$ représente la norme euclidienne dans $\mathbb{R}^n$.

Si les vecteurs colonnes de X sont linéairement indépendants, la solution est unique. Cette méthode conduit à un prédicteur $\hat{Y} = X\beta_{OLS}$ incluant toutes les variables. Si on suppose que seulement quelques variables ont vraiment un impact sur Y, alors le fait de toutes les considérer ajoute du bruit dans l'estimation des coefficients, ce qui contribue à diminuer le pouvoir prédictif du modèle à cause du sur-apprentissage. L'estimateur des moindres carrés a aussi un autre inconvénient de taille : il ne permet pas de résoudre un problème en grande dimension : lorsque $p > n$, les colonnes de la matrice X sont nécessairement linéairement dépendantes.

1.3.2 Les problèmes de grande dimension

Les problèmes en grande dimension sont des problèmes où le nombre de variables p est plus grand voire beaucoup plus grand que le nombre d'observations n, comme c'est souvent le cas avec des données transcriptomiques. Les problèmes en grande dimension sont insolubles en l'état puisque les analyses classiques -comme les moindres carrés- nécessitent un plus grand nombre d'observations que de variables explicatives. En effet, si on possède n observations, l'espace de travail est alors $\mathbb{R}^n$ et il est donc impossible d'estimer p coefficients si $p > n$.

Des hypothèses sont donc nécessaires pour résoudre les problèmes en grande dimension.

On suppose généralement que seule une petite portion des p variables porte le signal et on cherche à retrouver ce signal, c'est une condition de parcimonie ("sparsity" pour les anglo-saxons). Cette portion doit être inférieure en nombre à n, afin de ne pas retomber sur un problème de grande dimension et avoir des problèmes d'identifiabilité.

Si l'on note k le cardinal du support du vecteur des paramètres β, Verzelen (2012) montre que l'estimation de $X\beta$, ainsi que celle du support de β, est quasiment impossible lorsque $k\ ln\left(\frac{p}{k}\right)$ est grand devant n, appelé un cas de très haute dimension (ultra-high dimension). D'après l'expérience empirique de Verzelen (2012), la très haute dimension est définie par les cas vérifiant $\frac{k}{n}\ ln\left(\frac{p}{k}\right) > \frac{1}{2}$. Remarquons que cette condition est assez restrictive en pratique, pour $n = 50$ observations et $p = 500$ variables, une valeur de k égale à 6 nous place dans un cas de très grande dimension ($\frac{k}{n}\ ln\left(\frac{p}{k}\right) = 0.53$). Il est à noter que ce petit exemple n'est pas fortement éloigné de la réalité des données biologiques.

Nos données métabolomiques ne rentrent pas dans le cadre d'un problème de grande dimension au premier abord, en effet $n = 506 > 375 = p$. Néanmoins, le nombre de variables p peut très vite augmenter si l'on considère des interactions entre la race de l'individu et les données métabolomiques par exemple, c'est-à-dire si l'on suppose que les métabolites ont un effet différent suivant la race de l'animal. On considère alors comme matrice X une matrice contenant beaucoup plus de colonnes (au minimum le produit de p par le nombre de races) et le problème devient alors un problème de grande dimension insoluble avec l'estimateur des moindres carrés, mais pas avec la méthode Lasso.

1.3.3 La méthode Lasso

La méthode Lasso (Least Absolute Shrinkage and Selection Operator) est une pénalisation ℓ^1 des moindres carrés :

$$\beta^{\lambda}_{Lasso} = \underset{\beta\in\mathbb{R}^p}{Argmin} \left\{ ||Y - X\beta||_2^2 + \lambda \sum_{i=1}^{p} |\beta_j| \right\}, \lambda \geq 0. \tag{3}$$

L'ajout de la pénalité ℓ^1 sur le vecteur β a pour conséquence directe de mettre exactement à 0 certains coefficients de β^{λ}_{Lasso}, ce qui signifie que les variables correspondantes sont alors considérées comme non pertinentes, ou n'ayant aucune relation avec le phénotype Y. L'ensemble des variables mises à zéro dépend de la valeur de la pénalité ℓ^1 : pour une très forte pénalité -et donc une très grande valeur de λ- il ne reste aucune variable considérée comme pertinente, lorsque la pénalité diminue le nombre de variables pertinentes augmente, jusqu'à atteindre le modèle maximal pour une pénalité nulle (qui est alors un moindre carré ordinaire). Le choix de la pénalité est donc crucial, et de nombreuses techniques existent afin de faire un choix 'optimal'. Les principales techniques utilisées en pratique sont la validation croisée et la méthode BIC (Bayesian Information Criterion, Schwarz (1978)). Ces deux approches seront détaillées ultérieurement.

La méthode Lasso a longuement été étudiée et de nombreux résultats théoriques sont disponibles. Notamment, le Lasso est puissant sous la condition forte d'irreprésentabilité ou 'strong irrepresentable condition' de Zhao and Yu (2006). Si on définit J comme le support de β : $J = \{j, \beta_j \neq 0\}$, et notant X_J la sous-matrice de X construite à partir des colonnes J et X_{-J} la sous-matrice constituée des colonnes restantes, on peut écrire la 'strong irrepresentable condition' comme suit : $\exists \eta > 0$ tel que toutes les coordonnées du vecteur $\frac{1}{n} X'_{-J} X_J \left(\frac{1}{n} X'_J X_J \right)^{-1} \text{sign}(\beta_J)$ sont en valeur absolue majorée par $1 - \eta$, où le signe d'un vecteur $\beta \in \mathbb{R}^p$ est défini par :

$$\text{sign}(\beta) = (\text{sign}(\beta_1), \ldots, \text{sign}(\beta_p))' \text{ avec pour tout } j \in \{1, \ldots, p\}, \text{sign}(\beta_j) = \begin{cases} 1, \text{ si } \beta_j > 0 \\ 0, \text{ si } \beta_j = 0 \\ -1, \text{ si } \beta_j < 0 \end{cases}.$$

Sous cette condition ainsi qu'une condition sur le comportement de la pénalité λ, Zhao and Yu (2006) montrent que le Lasso est signe-consistent, c'est-à-dire que l'estimateur β^{λ}_{lasso} possède asymptotiquement les mêmes signes que le vrai paramètre β.

La 'strong irrepresentable condition' est en fait une condition sur le design de la matrice X, qui contraint les variables importantes (dans X_J) à ne pas être trop corrélées aux variables non pertinentes (dans X_{-J}). D'autres résultats théoriques sur le Lasso sont disponibles, notamment dans Wainwright (2009); Bunea et al. (2007); Zhang and Hunag (2008).

Obtenir les coefficients du Lasso en résolvant (3) étant un problème d'optimisation convexe, de nombreux algorithmes efficaces convergent rapidement vers la solution. Le plus connu est sans doute l'algorithme LARS (Least Angle Regression Stepwise, Efron et al. (2004)) qui fournit toutes les solutions du Lasso, c'est-à-dire l'ensemble des solutions de (3) pour une grande plage de pénalités λ - c'est le chemin de régularisation du Lasso-. Un exemple de chemin de régularisation est donné en Figure 2, provenant d'une analyse du jeu de données du cancer de la prostate fourni dans le package 'lasso2' du logiciel R, contenant 97 individus et 9 variables explicatives ($n = 97, p = 9$). Cette figure confirme que pour une forte pénalité aucune variable n'est sélectionnée (donc tous les coefficients sont à zéro), et quand la pénalité diminue les variables apparaissent jusqu'à ce que tous les paramètres soient estimés à l'aide d'un estimateur des moindres carrés classique pour une pénalité nulle.

L'utilisation en pratique de la méthode Lasso nécessite un choix approprié du paramètre de régularisation. Par ailleurs, le Lasso se montre peu stable et très dépendent des données : des petits perturbations dans les données peuvent impliquer de grands changements dans les résultats. Pour compenser ce problème, des extensions du Lasso ont été proposées, notamment la méthode Bolasso développée par Bach (2009) et l'adaptive Lasso introduit par Zou (2006), que nous allons détailler.

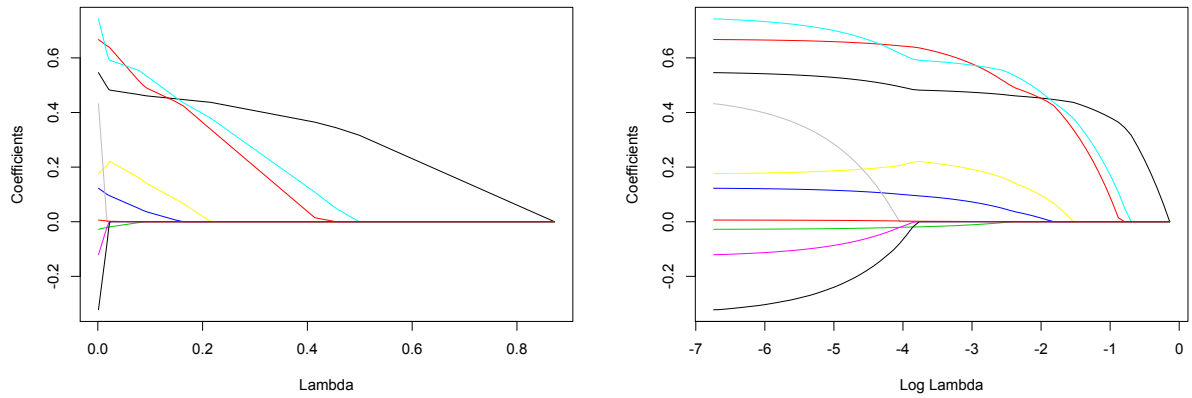


FIGURE 2 – Chemin de régularisation du Lasso pour les données Prostate du package 'lasso2'.

1.3.4 Quelques extensions de la méthode Lasso

Le Bolasso Afin de stabiliser l'ensemble des variables sélectionnées par la méthode Lasso, il est naturel d'introduire une procédure qui s'appuie sur le bootstrap. Ceci a été développé par Bach (2009) sous la terminologie de Bolasso. Le Bolasso est donc une version bootstrap du Lasso, ce qui en fait une version plus stable mais qui nécessite aussi le choix d'une pénalité. Le fonctionnement est simple : plusieurs échantillons bootstrap sont construits à partir du jeu de données, et le Lasso est appliqué sur chacun d'eux. Un ensemble de variables est sélectionné par le Lasso sur chaque échantillon bootstrap, l'intersection de tous ces ensembles constitue l'ensemble des variables considérées comme pertinentes pour le Bolasso. Le vecteur des coefficients β est ensuite estimé à partir d'une simple régression linéaire du vecteur des observations Y sur les données considérées comme pertinentes $X_{\hat{J}}$, si on note $\hat{J}$ l'estimation du support de β par le Bolasso.
Bach (2009) propose deux variantes de cette méthode, un "random pair bootstrap" et un "bootstrapping residuals". Le premier est un bootstrap sur les observations, il consiste à tirer aléatoirement avec remise n couples (X^i, Y_i), où X^i correspond aux p données de l'individu i, $1 \leq i \leq n$. On obtient ainsi une nouvelle matrice de données $\tilde{X}$ et un nouveau vecteur d'observations $\tilde{Y}$, le Lasso est appliqué sur chaque nouvel échantillon bootstrap $(\tilde{X}, \tilde{Y})$.
Le second est un bootstrap sur les résidus. Notons $\hat{\beta}$ une estimation de β et $\tilde{\epsilon} = Y - X\hat{\beta}$ le vecteur des résidus estimés. On note $\hat{\epsilon}$ les résidus centrés $\hat{\epsilon} = \tilde{\epsilon} - \frac{1}{n}\sum_{i=1}^{n} \tilde{\epsilon}_k$. Le bootstrap sur les résidus consiste à construire les échantillons suivants : $Y_{i^*} = X\hat{\beta} + \hat{\epsilon}_{i^*}$ où i^* est tiré aléatoirement avec remise dans $\{1, \ldots, n\}$. La matrice X est donc inchangée, et on construit de nouvelles observations Y^* à partir d'une première estimation du vecteur β.

Sauf mention contraire, quand le Bolasso sera mentionné, il sera fait référence au bootstrap sur les résidus puisque c'est celui qui donne les meilleurs résultats en termes de sélection de variables en grande dimension d'après Bach (2009).

Adaptive Lasso L'adaptive Lasso (Zou, 2006) est actuellement une variante de choix. Notons $(w_1, \dots, w_p)$ une suite de valeurs strictement positives, la solution de l'adaptive Lasso est définie comme suit :

$$\beta^{\lambda}_{adLasso} = \underset{\beta \in \mathbb{R}^p}{Argmin} \left\{ ||Y - X\beta||_2^2 + \lambda \sum_{i=1}^{p} w_j |\beta_j| \right\}, \lambda \geq 0. \tag{4}$$

Cette méthode rajoute un paramètre par rapport au Lasso classique qui est le vecteur des poids $(w_1, \dots, w_p)$. Ces poids sont en général définis à partir de l'estimateur des moindres carrés : $w = 1/|\beta_{OLS}|$. Dans un problème de grande dimension, les moindres carrés ne pouvant pas être calculés à partir du modèle (1), les poids sont donc définis de manière analogue en calculant une estimation de β_j pour tout $1 \leq j \leq p$ par moindres carrés dans le modèle $Y = X_j\beta_j + \epsilon_j$.
La résolution de (4) restant un problème convexe, les algorithmes développés pour la résolution du Lasso peuvent s'adapter en considérant des pénalités différentes pour chaque coefficient β_j. Commentons le comportement de l'adaptive Lasso dans deux cas extrêmes. Si on considère le poids w_j comme infini, alors cela revient à exclure du modèle la variable X_j correspondante. Au contraire, si le poids w_j est nul, alors la variable correspondante est incluse par défaut dans le modèle. Les poids initiaux ont donc un impact non négligeable dans la solution de (4). Nous verrons ultérieurement que définir les poids initiaux à l'aide des moindres carrés n'est pas toujours la bonne solution.

Les méthodes Bolasso et adaptive Lasso souffrent du même problème que le Lasso original, à savoir le choix de la pénalité qui conditionne fortement les résultats.

1.3.5 Le choix de la pénalité

Nous allons détailler deux techniques -validation croisée et critère BIC- pour choisir la pénalité du Lasso ou de ses extensions.

Validation croisée La validation croisée ou "cross-validation " est une méthode fondée sur une technique d'échantillonnage. La méthode consiste à se fixer un entier $k \in \{2, \dots, n\}$ puis à diviser les données en k parts sensiblement de même taille. On sélectionne un des k échantillons comme ensemble de validation et les $(k-1)$ autres échantillons constituent alors l'ensemble d'apprentissage. Le modèle est bâti sur l'ensemble d'apprentissage et testé sur l'ensemble de validation en calculant une erreur quadratique moyenne. On répète l'opération k fois pour que chaque paquet serve une seule fois d'échantillon de validation, et on moyenne ensuite les k erreurs quadratiques moyennes afin d'obtenir une

estimation de l'erreur de prédiction de la méthode utilisée pour bâtir le modèle.
La méthode "leave-one-out" est un cas particulier de la "k cross-validation" lorsqu'on prend $k = n$. Cette méthode a l'avantage de ne pas dépendre de la manière dont sont construits les paquets (contrairement à la "k cross-validation"), mais elle est moins rapide (le modèle est appris n fois au lieu de k fois).

La technique de validation croisée est donc appliquée pour différentes valeurs de λ et la valeur qui minimise l'erreur quadratique moyenne est considérée comme la pénalité optimale.

Critère BIC et EBIC Le 'Bayesian Information Criterion' (Schwarz, 1978) est un critère de vraisemblance pénalisé qui permet de faire de la sélection de modèle. Ce critère provient d'une approximation asymptotique d'un critère de choix de modèle bayésien : on cherche le modèle ayant l'*a posteriori* le plus probable en ayant considéré un *a priori* uniforme pour tous les modèles. La log-vraisemblance L du modèle (1) est donnée par :

$$-2L(\beta;\sigma) = n\ ln(2\pi) + n\ ln(\sigma^2) + ||Y - X\beta||^2/\sigma^2. \tag{5}$$

Soit $(\hat{\beta}, \hat{\sigma})$ l'estimateur du maximum de vraisemblance de (β, σ) dans le modèle (1). Alors, $-2L(\hat{\beta};\hat{\sigma}) = n\ ln(2\pi) + n\ ln(||Y - X\hat{\beta}||^2/n) + n$. Si on note $(\hat{\beta}_S, \hat{\sigma}_S)$ l'estimateur par maximum de vraisemblance de (β, σ) sur un modèle S de dimension k, le critère BIC associé à ce modèle est, par définition,

$$-2L(\hat{\beta}_S;\hat{\sigma}_S) + k\ ln(n). \tag{6}$$

Le modèle ayant le critère BIC le plus faible parmi une collection de modèles est sélectionné. Cette technique s'applique pour choisir le paramètre λ du Lasso (3) ou de ses extensions en cherchant à minimiser le critère suivant pour une plage de différentes valeurs de λ :

$$n\ ln(||Y - X\beta^\lambda||^2/n) + |\beta^\lambda|_0\ ln(n), \tag{7}$$

où β^λ est l'estimateur obtenu par la méthode considérée (donc β^λ_{Lasso}, $\beta^\lambda_{adLasso}$ ou autres) et $|\beta^\lambda|_0$ est le nombre de composantes non nulles du vecteur β^λ.

Le critère BIC est consistent en sélection de modèle (Rao and Wu, 1989) lorsque n tend vers l'infini et p est fixé. Cependant, il n'est pas conçu pour la sélection lorsque le nombre de paramètres est très grand. Chen and Chen (2008) ont donc considéré un *a priori* différent pour chaque modèle S et non un *a priori* uniforme pour tous les modèles comme c'est le cas pour le critère BIC. Cet *a priori* dépend du nombre de modèles ayant la même dimension que S. Chen and Chen (2008) montrent que le critère EBIC ainsi obtenu est consistent pour une valeur de p polynomiale en n, sous une simple condition d'identifiabilité.

Parallèlement à l'utilisation de critères pénalisés, des méthodes de sélection de variables basées sur des procédures de tests multiples, ne nécessitant pas de pénalité, ont été développées, notamment la procédure FDR.

1.3.6 La procédure FDR (False Discovery Rate)

La procédure FDR, basée sur des tests multiples, est largement employée en biologie pour découvrir des gènes différentiels, par exemple entre deux conditions A et B. Admettons qu'il y ait p gènes, l'analyse consiste à tester indépendamment chacune des p hypothèses nulles "le gène i n'est pas différentiellement exprimé entre la condition A et la condition B" pour tout $1 \leq i \leq p$. Chaque test est réalisé avec une erreur de première espèce fixée au préalable. Le risque de première espèce d'un test d'hypothèse représente la probabilité de rejeter à tort l'hypothèse nulle alors qu'elle est vraie, le risque de seconde espèce étant la probabilité d'accepter l'hypothèse nulle alors qu'elle est fausse.
Fixons le risque de première espèce $\alpha = 0.05$. Si on fait un seul test d'hypothèse, la probabilité de rejeter l'hypothèse à tort est de $1-(1-\alpha) = 5\%$. Si on fait deux tests indépendants, la probabilité de rejeter au moins une hypothèse à tort est de $1-(1-\alpha)^2 = 9.75\%$. Pour 100 tests indépendants, on a une probabilité de 99.4% de rejeter au moins une hypothèse à tort.
Sur ce petit exemple, il paraît clair que conduire une telle analyse pour trouver des gènes différentiels donnerait nombre de faux-positifs (gènes considérés comme différentiellement exprimés à tort). C'est pourquoi plusieurs méthodes ont vu le jour afin de prendre en compte les tests multiples. Elles sont basées sur un contrôle des faux-positifs, que ce soit un contrôle global par le contrôle du FWER (Family Wise Error Rate) ou un contrôle de la proportion de faux-positifs par le contrôle du FDR (False Discovery Rate). La première méthode contrôle la probabilité que le nombre d'hypothèses rejetées à tort V soit supérieur à 1 ; la seconde contrôle la proportion de faux positifs : si on note R le nombre total d'hypothèses rejetées, alors contrôler le taux de faux-positifs au niveau α signifie : $\mathbb{E}(Q) \leq \alpha$ où $Q = \begin{cases} V/R \text{ si } R > 0 \\ 0 \text{ sinon} \end{cases}$. Les méthodes les plus connues sont la méthode Bonferonni (contrôle du FWER) et les méthodes de Benjamini and Hochberg (1995) et Benjamini and Yekutieli (2001) (contrôle du FDR).

Ces méthodes de contrôle de faux-positifs ont été appliquées en sélection de variables par Bunea et al. (2006). Les hypothèses nulles considérées sont les suivantes :

$$H_i : \beta_i = 0, \qquad i = 1, \ldots, p.$$

Pour tester ces hypothèses, on estime le vecteur β par $\hat{\beta}$ dans le modèle (1) à l'aide de l'estimateur des moindres carrés (2). Pour tout $1 \leq j \leq p$, l'écart type $se(\hat{\beta}_j)$ est calculé ainsi que la statistique de Student $t_j = \hat{\beta}_j / se(\hat{\beta}_j)$, les p-valeurs sont obtenues par $\pi_j = 2\{1 - \Phi(|t_j|)\}$ où Φ est la fonction de répartition de la loi normale centrée réduite.

Les méthodes décrites précédemment peuvent être appliquées : La méthode Bonferonni estime le support J de β par $\hat{J} = \{i : \pi_i \leq \alpha/p\}$ tandis que la procédure FDR utilise la méthode de Benjamini and Yekutieli (2001) et est appliquée comme suit :
On ordonne les p-valeurs $\pi_{(1)} \leq \cdots \leq \pi_{(p)}$ et on définit $k = \max \left\{ i : \pi_{(i)} \leq \frac{i}{p} \frac{\alpha}{\sum_{j=1}^{p} j^{-1}} \right\}$.

Si un tel k existe, on estime J par $\hat{J} = \{(1), \ldots, (k)\}$, sinon $\hat{J} = \emptyset$.

La procédure FDR contrôle le taux de faux-positifs et la consistance a été montrée par Bunea et al. (2006) : $\lim_{n\to\infty} \mathbb{P}(\hat{J} = J) = 1$ lorsque p tend vers l'infini avec n mais pas plus vite que $\sqrt{n}$ et sous certaines conditions sur la matrice X. La procédure est donc consistante lorsque $p < \sqrt{n}$, soit pour des problèmes de petite dimension. Néanmoins, cette procédure est couramment utilisée en grande dimension ($p > n$), chaque coefficient β_j est estimé dans un modèle linéaire $Y = X_j\beta_j + \epsilon_j$ et le reste de la méthode est appliqué. Il n'existe néanmoins pas de justification de la procédure dans ce cadre.

Le modèle linéaire est le modèle le plus classique dans lequel se placer pour analyser des données. Cependant, comme nous l'avons signalé dans la description des données dont nous disposons (cf. Section 1.2), certains éléments additionnels peuvent être pris en compte comme les liens de parentés entre les individus ou l'environnement dans lequel ils ont grandi. Ces variables n'étant pas des variables d'intérêt en elles-même mais plutôt des variables à considérer comme du bruit, on se focalise dans la suite sur les modèles linéaires mixtes et la sélection de variables dans ces modèles.

1.4 La sélection de variables dans un modèle linéaire mixte

Dans un modèle linéaire classique, les observations sont supposées indépendantes et généralement identiquement distribuées. Lorsqu'une structure sur les données est disponible, comme une structure familiale, ces hypothèses ne sont plus adaptées. Cette structure familiale peut être prise en compte dans un modèle linéaire mixte en considérant le facteur famille comme un effet aléatoire. Les effets aléatoires sont modélisés par des variables gaussiennes, et on ne s'intéresse qu'à la variance de ces effets aléatoires ; les variables non aléatoires (comme les métabolites) sont appelées des effets fixes. Pour prendre en compte cette structure des observations, le modèle mixte considère une matrice de variance-covariance des observations V non plus diagonale mais diagonale par blocs ; les blocs étant construits à l'aide de la structure considérée.
Le modèle linéaire mixte a reçu une attention considérable pour l'estimation des composantes de la variance. Deux méthodes récentes sont couramment utilisées, une estimation par maximum de vraisemblance (ML) (Henderson, 1973, 1953) et une estimation par maximum de vraisemblance restreint (REML) qui prend en compte la perte de degrés de liberté due à l'estimation des effets fixes du modèle (Patterson and Thompson, 1971; Harville, 1977; Henderson, 1984; Foulley et al., 2002).

Détaillons le modèle linéaire mixte avant d'expliciter l'état de l'art sur la sélection d'effets fixes dans le modèle linéaire mixte.

1.4.1 Le modèle linéaire mixte

Le modèle linéaire mixte se décrit dans le modèle marginal :

$$y = X\beta + \zeta, \zeta \sim \mathcal{N}(0, V), \tag{8}$$

où y est le vecteur des données observées de longueur n, $X = (X_1, \ldots, X_p)$ est la matrice des p effets fixes et $\beta = (\beta_1, \ldots, \beta_p)$ est un vecteur inconnu de $\mathbb{R}^p$. La matrice V est une matrice diagonale par bloc où les blocs représentent la structure des observations.

L'estimation des paramètres à l'aide des approches ML et REML prend en compte la totalité des effets fixes (les β_j), or comme on l'a vu dans le modèle linéaire classique, cette hypothèse peut entraîner une estimation fausse des paramètres d'intérêt, en plus d'une impossibilité de les estimer en grande dimension ($p > n$). La sélection de variables, ou sélection d'effets fixes, apparaît comme nécessaire dans ce contexte. Cependant, peu de méthodes existantes répondent à ce problème. Bondell et al. (2010) et Ibrahim et al. (2011) ont introduit un critère de vraisemblance pénalisée qui permet de faire de la sélection à la fois sur les effets fixes et sur les effets aléatoires. Cependant, leurs simulations ne concernent que la petite dimension. Seuls Schelldorfer et al. (2011) ont vraiment étudié la question dans un contexte de grande dimension, grâce à la méthode 'lmmLasso'. La sélection d'effet fixes dans le modèle linéaire mixte peut aussi être envisagé à travers le domaine de la sélection de modèles (de manière similaire au problème de sélection de variables dans le modèle linéaire), notamment à l'aide de critère pénalisé (Lavergne et al., 2008).

1.4.2 La méthode lmmLasso

La méthode lmmLasso permet de faire de la sélection d'effets fixes dans un modèle linéaire mixte ; elle repose sur une pénalisation ℓ^1 de la vraisemblance du modèle marginal (8). La log-vraisemblance de (8) étant

$$L(\beta, V) = \frac{1}{2}\left\{ln(2\pi) + ln\ |V| + (y - X\beta)'V^{-1}(y - X\beta)\right\}, \tag{9}$$

où $|V|$ est le déterminant de la matrice V, la fonction objectif à minimiser en les paramètres du modèle, que sont β et V, est :

$$Q_\lambda(\beta, V) = \frac{1}{2}ln\ |V| + \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta) + \lambda\sum_{i=j}^{p}|\beta_j|, \tag{10}$$

où λ est un paramètre de régularisation positif. Cette fonction objectif étant non convexe, les auteurs ont proposé un algorithme de descente de gradient qui converge vers un minimum local de la fonction objectif. Leur algorithme repose sur l'inversion de la matrice de variance V, ce qui peut s'avérer coûteux en temps de calcul si le nombre total d'observations n est grand (V est de taille $n \times n$).
Schelldorfer et al. (2011) ont aussi proposé une extension du lmmLasso, le lmmadLasso. La fonction objectif (10) est modifiée pour prendre en compte une famille de poids positifs $w_1, \ldots, w_p$:

$$Q_\lambda^{w_1,\ldots,w_p}(\beta, V) = \frac{1}{2}ln\ |V| + \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta) + \lambda\sum_{i=j}^{p}w_j|\beta_j|, \tag{11}$$

La méthode lmmLasso est consistante sous certaines conditions sur le signal et sur les matrices X et Z ; des inégalités oracle sont démontrées pour la méthode lmmadLasso. Ces deux méthodes sont performantes sur les simulations présentes dans l'article des auteurs. Néanmoins, ces deux méthodes sont relativement longues en temps de calcul lorsque le nombre d'observations n est grand, ce qui est le cas pour les données métabolomiques qui portent sur $n = 506$ observations.

1.5 Plan du manuscrit

La suite de ce manuscrit se décompose en trois parties. Une partie se focalise sur le premier problème soulevé dans cette introduction qui est la prédiction de phénotypes d'intérêt à l'aide de données métabolomiques (Partie 2). Pour se faire, le mode d'obtention des données métabolomiques est précisé et la combinaison d'une transformée en ondelettes et d'une méthode de sélection de variables (la méthode Lasso) est proposée. On montre notamment que cette combinaison permet de mieux prédire certains phénotypes d'intérêt qu'une simple analyse Lasso. Le point important de cette partie est de montrer que les données métabolomiques -et donc une simple prise de sang- sont capables de prédire certains phénotypes d'intérêt comme le taux de muscle avec des taux d'erreurs très convenables, malgré l'espace temporel entre le moment de la prise de sang et le moment de la mesure des phénotypes. Cette étude laisse envisager un réel pouvoir prédictif du métabolome en temps réel. Cet article est accepté pour publication à *Journal of Animal Science*.

La Partie 3 sera consacrée à de nouvelles méthodes de sélection de variables dans un modèle linéaire développées au cours de ce travail de thèse. Deux méthodes ont vu le jour, elles sont toutes deux des procédures séquentielles de tests multiples basées sur une procédure développée par Baraud et al. (2003). Une méthode concerne la sélection ordonnée, et une autre la sélection non ordonnée. Elles sont toutes deux puissantes sous certaines conditions sur le signal et fonctionnent en grande dimension ($p > n$). Les résultats de simulations de ces deux méthodes sont très bons, et ceux de la méthode pour la sélection non ordonnée surpassent les méthodes classiques, surtout dans des modèles de grande dimension. A noter que la méthode développée pour la sélection ordonnée n'est pas comparable aux méthodes classiques car un a priori sur l'importance des variables est connu. Ce travail est soumis.

La dernière partie (Partie 4) présente une nouvelle méthode de sélection des effets fixes dans un modèle linéaire mixte qui fonctionne en grande dimension ($p > n$) tout en supposant peu d'effets aléatoires. Cette méthode est similaire à la méthode lmmLasso (cf. Section 1.4.2) ; les résultats de simulations sont d'ailleurs très similaires, mais la méthode développée est beaucoup plus rapide puisqu'elle ne nécessite pas l'inversion d'une matrice $n \times n$. L'algorithme présenté pour résoudre le problème d'optimisation non convexe de

notre méthode est un multicycle ECM (Foulley, 1997; McLachlan and Krishnan, 2008; Meng and Rubin, 1993). Il sera détaillé en Section 4.2. Cet algorithme permet l'utilisation de n'importe quelle méthode de sélection de variable développée pour le modèle linéaire classique. On peut donc combiner l'algorithme à l'adaptive Lasso ou à la procédure de tests multiples présentée dans la Partie 3. Cependant, seule une méthode qui optimise un critère (comme le Lasso (3) ou l'adaptive Lasso (4)) permet d'obtenir des résultats de convergence de l'algorithme. Cet algorithme permet aussi de faire une sélection sur les effets aléatoires. Ce travail sera prochainement soumis.

2 Prédiction phénotypique à l'aide de données métabolomiques

2.1 Contexte

Obtenir une bonne prédiction d'un phénotype d'intérêt économique dans l'espèce porcine à partir d'une simple prise de sang peut avoir des conséquences importantes dans le monde de l'exploitation. En effet, une prise de sang n'est pas une opération invasive et si elle suffit à faire aussi bien dans la détermination de caractères importants qu'un abattage, alors les éleveurs ont tout à gagner dans la généralisation de cette technique. L'article présenté dans la section suivante se place donc dans ce contexte de prédiction de phénotypes à l'aide de données métabolomiques. Les 27 phénotypes explicités dans la Table 1 ont été étudiés individuellement. A noter qu'ils ne sont pas tous d'intérêt économique majeur. Les données des 506 individus à notre disposition proviennent de 3 races et de 8 bandes, le plan d'expérience est détaillé dans la Table 2. Ce plan d'expérience est très déséquilibré ce qui perturbe l'estimation des paramètres. Il faut donc relativiser les résultats obtenus pour la prédiction des phénotypes en considérant qu'un plan d'expérience plus équilibré pourrait permettre l'obtention de meilleurs résultats. Il est toutefois à noter que les données sont recueillies sur le terrain et que tous les paramètres ne sont pas maîtrisables.

	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8
Large White Femelle	42	45	54	13	16	20	8	0
Landrace	22	39	51	0	21	28	26	0
Piétrain	0	37	29	5	0	33	0	17

TABLE 2 – **Répartition des animaux dans chaque groupe, race et bande**

On suppose une relation linéaire entre chaque phénotype et les données métabolomiques, on se place donc dans le modèle (1). Les données métabolomiques sont centrées et réduites ($\mathbb{E}(X_i) = 0$ et $\sum_j X_{i,j}^2 = n, \forall 1 \leq i \leq n$). Afin de prendre en compte le plan d'expérience, l'article présenté dans la section suivante se focalise sur l'étude des trois modèles suivants :

$$\text{phénotype} = \text{intercept} + \text{metab} + \text{bruit} \tag{12a}$$

$$\text{phénotype} = \text{intercept} + \text{race} + \text{metab} + \text{metab*race} + \text{bruit} \tag{12b}$$

$$\text{phénotype} = \text{intercept} + \text{race} + \text{bande} + \text{metab} + \text{metab*race} + \text{metab*bande} + \text{bruit}, \tag{12c}$$

où metab correspond aux données métabolomiques et metab*race signifie qu'on considère un effet d'interaction entre les métabolites et la race des animaux. Dans le modèle (12b),

on considère que les données métabolomiques ont à la fois un effet global, mais aussi un effet dépendant de la race. Le modèle (12c) fait de même pour la race et la bande. Ces trois modèles (12) ne correspondent pas tous à des problèmes de grande dimension. En effet, le premier modèle (12a) est un problème de "petite dimension" puisque les données sont constituées de $n = 506$ individus pour $p = 375$ paramètres métabolomiques et 1 paramètre de plus pour l'intercept. Par contre l'estimation des paramètres dans les deux autres modèles est un problème de grande dimension. En effet le modèle (12b) contient 1504 paramètres et le modèle (12c) en contient 4512.
Les résultats de l'analyse de chacun des 27 phénotypes sur chacun des 3 modèles constituent un article accepté pour publication à "Journal of Animal Science" qui est présenté dans la section suivante.

2.2 Article - Prédiction de phénotypes à partir du métabolome

Résumé La prédiction de phénotype est un défi statistique et biologique, à la fois en médecine (prédire une maladie) et en production animale (prédire la valeur économique de la carcasse d'un jeune animal). Le but de ce travail était de quantifier le pouvoir prédictif des profils métabolomiques pour des phénotypes de production à partir d'une simple prise de sang sur le porc en croissance. Différentes méthodes statistiques ont été comparées sur la base de la validation-croisée : les données brutes vs une transformée du signal (les ondelettes), avec une seule méthode de sélection de variables. Les meilleurs résultats en terme d'erreur de prédiction ont été obtenus quand les données furent transformées en ondelettes dans la base de Daubechies.
Les phénotypes considérés comme de bons indicateurs de la qualité de la viande n'ont pas été particulièrement bien prédits puisque la prise de sang était relativement espacée de l'abattage, or l'abattage est connu pour avoir une forte influence sur ces paramètres. Néanmoins, des phénotypes d'intérêt économique comme le taux de muscle (LMP) ou la consommation moyenne journalière (DFI) ont été bien prédits à partir des données métabolomiques ($R^2 = 0.7$).

Phenotypic Prediction based on Metabolomic Data on the Growing Pig from three main European Breeds

F. Rohart[1,2], A. Paris[3], B. Laurent[2], C. Canlet[4], J. Molina[4], M.J. Mercat[5], T. Tribout[6], N. Muller[7], N. Iannuccelli[1], N. Villa-Vialaneix[8], L. Liaubet[1], D. Milan[1] and M. San Cristobal[1]

[1] INRA, UMR444 Laboratoire de Génétique Cellulaire, F-31326 Castanet Tolosan, France
[2] INSA, Département de Génie Mathématiques, and Institut de Mathématiques, Université de Toulouse (UMR 5219), F-31077 Toulouse, France
[3] INRA, Met@risk, F-75231 Paris Cedex 05, France
[4] INRA, UMR 1331 Toxalim (Research Centre in Food Toxicology), INRA/INP/UPS, F-31027 Toulouse, France
[5] BIOPORC, 75595 PARIS Cedex 12
[6] INRA GABI, F-78351 Jouy-en-Josas cedex, France
[7] INRA UE450 Testage - Porcs, F-35653 Le Rheu, France
[8] SAMM, Université Paris 1, 75013 Paris, France

The authors thank the animal and DNA providers (BIOPORC) and French ANR for funding the DéLiSus project (ANR-07-GANI-001). F.R. acknowledges financial support from Région Midi-Pyrénées. Thanks to Hélène Gilbert for interesting discussions and Helen Munduteguy for the English revision.

Abstract

Predicting phenotypes is a statistical and biotechnical challenge, both in medicine (predicting an illness) and animal breeding (predicting the carcass economical value on a young living animal). High-throughput fine phenotyping is possible using metabolomics, which transcribes the global metabolic status of an individual, and is the closest to the terminal phenotype. The purpose of this work was to quantify the prediction power (in the statistical sense) of metabolomic profiles for commonly used production phenotypes from a single blood sample on the growing pig. Several statistical approaches were investigated and compared on the basis of cross validation: raw data vs. signal preprocessing (wavelet transform), with a single feature selection method. The best results in terms of prediction accuracy were obtained when data was preprocessed using wavelet transforms on the Daubechies basis. The phenotypes related to meat quality were not particularly well predicted since the blood sample is taken some time prior to slaughter, and slaughter is known to have a strong influence on these traits. In contrast, phenotypes of potential economic interest, e.g. lean meat percentage and daily feed intake, were well predicted using metabolomic data ($R2 = 0.7$).

Key Words: metabolome, phenotypic prediction, variable selection, wavelet transform, pig

1 Introduction

The accurate and competitive prediction of production phenotypes may open new perspectives for livestock selection. For instance, phenotypes of interest could be those which are of considerable economic importance, and have top priority in selection objectives, but are too expensive to measure routinely or for which measurement is too invasive. Metabolomics is a relatively cheap and easy way to predict (reviewed by Rochfort, 2005) or discover promising biomarkers (Zhang et al., 2011). Recently, this approach has been successfully used in the pig to compare highly phenotypically differentiated breeds (D'Alessandro et al., 2011; He et al., 2012), but not to predict commercially important phenotypes in various breed $\times$ gender determined conditions involving European pig breeds.
The present work was motivated by the hypothesis that the blood metabolome could predict some production phenotypes, prediction being meant in the statistical sense. The rationale is that the blood metabolism reflects the general physiological state of the animal which is resulting from the functional metabolic state of the different tissues, since blood carries a lot of metabolites, hormones, etc, between them. The objective of this paper is to quantify on real data the power of prediction of several production phenotypes obtained by metabolomic data coming from a single blood sample. The chosen strategy is to evaluate the influence of external factors, namely breed and batch (that reflects micro-variations of the environment). Meanwhile, concurrent statistical tools will be also evaluated, in particular the signal pre-treatment step, and the final biological coherence of results will be discussed.

2 Materials and methods

2.1 Animal handling and zootechnical data

All procedures and facilities were approved by French veterinary services. A total of 506 animals from a Large White dam breed (LW), a Landrace dam breed (LR) and a Piétrain sire breed (PI) were considered in the analysis.
The animals (castrates in LW and LR, females in PI) were raised at the French central test Station in Le Rheu (France) in 2007 and 2008, in 8 different batches. The sampling design for breeds and batches is given in Table 1. Pigs were grouped in pens of 12 animals from the beginning of the test period ($\sim$ 10 weeks of age) until the day before slaughter, considered as the end of the test period ($\sim$ 110 kg live weight). They were given ad libitum access to water and to a standard pelleted diet formulated to contain 13.2 MJ digestible energy/kg and 164 g crude protein / kg feed. Pens were equipped with ACEMA 64 electronic feeders, allowing the recording of individual food consumption (Labroue et al., 1993). Animals were individually weighted at the beginning of the test period, at the

end of the test period (LWETP), and a last time before departure to the slaughterhouse (LWS) after at least 16 hours of fasting. The duration of the test period, LWETP and the individual feed consumption during the test period (FCTP) were used to calculate the average daily gain (ADG), the feed conversion ratio (FCR) and the daily feed intake (DFI) during the test period. Slaughters occurred at a given weight on a fixed day in the week in a commercial slaughterhouse (Cooperl-Hunaudaye, Montfort-sur-Meu, France). Carcass weight with and without the head (CW and CWwtH, respectively) and the weight of the right half-carcass (HCW) were recorded post-evisceration on the day of slaughter, and the dressing percentage (DP) was calculated as $CW \times 100/LWS$. The day after slaughter, the length of the carcass from the pubis to the atlas (Length), as well as the backfat thickness at the shoulder, last rib and hip joint at the sectioned edge of the carcass (BFsh, BFlr and BFhj, respectively) were recorded. The mean of these 3 fat measurements was calculated (mBF). The measurements used for carcass commercial grading, i.e. backfat thickness between the third and fourth lumbar vertebrae (G1) and between the third and fourth last ribs (G2), as well as loin eye depth between the third and fourth last ribs (M2), were performed using a "CGM" probe (Daumas et al., 1998) and were combined to estimate the commercial lean meat percentage (ComLMP). Finally, a standardized cutting procedure of the right half carcass was then performed, as described in Anonymous (1990), and ham, loin, backfat, shoulder and belly were weighed (hamW, loinW, bfW, shW, beW, respectively) and combined to obtain a second estimate of the lean meat percentage of the carcass (LMP; Metayer and Daumas, 1998). On the same day, several meat quality measurements were taken: the ultimate pH of the Semimembranous muscle (pH24), the color of the Gluteus superficialis muscle through the 3 coordinates (L*, a* and b* system) using a CR-300 Minolta Chromameter, and the water holding capacity of the Gluteus superficialis muscle (WHC). WHC, pH24 and L* were combined to compute a synthetic meat quality index (MQI) defined as a predictor of the technological yield of cured-cooked Paris ham processing, as described by the Institut Technique du Porc (1993). In total, 27 traits were recorded on the animals.

2.2 Metabolomic data

Blood samples were collected on sodium heparin once for every animal during the test period when animals displayed a weight of approx. 60 kg. Samples were immediately centrifuged at $2,500$ g for 15 min at $4°C$ to separate plasma from red cells and stored at -80°C until analysis.

Fingerprinting was performed by 1H NMR spectroscopy after a rapid sample preparation performed as follows: D2O ($500\mu l$) was added to plasma ($200\mu l$) and mixed, the sample was then centrifuged for 10 min at 3,000 g and the supernatant ($600\mu l$) was transferred to 5 mm NMR tubes for 1H NMR (Nuclear Magnetic Resonance) analysis.

All 1H NMR spectra were acquired on a Bruker Avance $DRX-600$ spectrometer (Bruker SA, Wissembourg, France) operating at 600.13 MHz for 1H resonance frequency, and equipped with a pulsed field gradients z system, an inverse 1H-13C-15N cryoprobe attached to a cryoplatform (the preamplifier cooling unit), and a temperature control unit maintaining the sample temperature at 300 ± 0.1 K.
The 1H NMR spectra of plasma samples were acquired at 300K using the Carr-Purcell-Meiboom-Gill (CPMG) spin-echo pulse sequence with presaturation with a total spin-echo delay ($2n\pi$) of 320 ms to attenuate broad signals from proteins and lipoproteins, which otherwise display a wide signal and hide the narrower signals of low molecular weight metabolites. The 1H signal was acquired by accumulating 128 transients over a 12-ppm spectral width, collecting $32,000$ data points. The interpulse delay of the CPMG sequence was set at 0.4 ms with n equal to 400 as defined in the following sequence: $[90-(\tau-180-\tau)n$ acquisition]. A 2-s relaxation delay was applied. The Fourier transform (FT) was calculated on $64,000$ points. All 1H NMR spectra were phased, and the baseline corrected. The 1H chemical shifts were calibrated on the resonance of lactate at 1.33 ppm. Then, serum spectra were data-reduced prior to statistical analysis using AMIX software (Analysis of Mixtures v 3.8) from Bruker Analytische Messtechnik (Rheinstetten, Germany). The spectral regionδ $0.5-10.0$ ppm was segmented into consecutive non-overlapping regions of 0.01 ppm (buckets) and normalized according to the total signal intensity in every spectrum. The region aroundδ 4.8 ppm corresponding to water resonance was excluded from the pattern recognition analysis to eliminate artifacts of residual water. Eight hundred and eleven quantitative variables were obtained for every spectrum and were processed by a multidimensional scaling-based procedure to select only informative metabolic variables. More precisely, the multidimensional scaling step which was repeatedly used ($n = 8$) to select fully informative variables was performed on the transposed matrix of data. Multidimensional scaling is a multidimensional statistical technique which corresponds here to a principal component analysis (PCA) of the matrix of distances between variables. Fully informative metabolic variables display a larger variance than baseline variables and therefore the distances between these two types of variables is larger than the distances between the sole baseline variables. Thus, at each selection step and for every variable, we calculated a distance between the origin and projection coordinates of the variable on the first factorial plan, and variables displaying the larger distances were subsequently selected. After 8 selection steps, only baseline relevant variables were remaining in the unselected dataset and were not included in the informative dataset on which further statistical analyses were achieved. Finally, each metabolomic profile or spectrum was observed on a discrete sampling grid of size $p = 375$ (number of buckets) as plotted in Figure 1.
Technical duplicates were performed on a limited number of animals, and showed a good adequacy between them (not shown), as expected. Since it was impossible to standardize feeding conditions in the farm, nor exact age, large samples within breeds were performed.

The result of a metabolomic experiment is a spectrum, in which some points are known to correspond to one or several metabolites, but not all. Identification of candidate informative metabolites (after the statistical treatment described below) was performed from known chemical shift references acquired on standard compounds and found in the literature or in a home-made reference databank. 2D homonuclear 1H-1H COSY (Correlation Spectroscopy) and 2D heteronuclear 1H-13C HSQC (heteronuclear single quantum coherence spectroscopy) NMR spectra were also registered for selected samples as an aid to spectral assignment. For COSY NMR spectra, a total of 32 transients were acquired into 1024 data points. A total of 256 increments were measured in F1 using a spectral width of 10 ppm and an acquisition time of 0.28 s was used. The data were weighted using a sine-bell function in the two dimensions prior to Fourier transformation. For HSQC NMR spectra, a relaxation delay of 2.5 s was used between pulses, and a refocusing delay equal to 1/41JC-H (1.78 ms) was employed. A total of 1024 data points with 64 scans per increment and 512 experiments were acquired with spectral widths of 10 ppm in F2 and 180 ppm in F1. The data were multiplied by a shifted Qsine-bell function prior to Fourier transformation.

2.3 Wavelet pre-processing

As proposed by Davies et al. (2007) and Xia et al. (2007), each metabolomics profile was written as the sum of weighted elementary functions, describing hierarchically the signal from a rough tendency to the finest details, in a finite number of resolution levels. Here, each one of the 506 spectra was decomposed onto a Haar basis (elementary step functions). The corresponding wavelet coefficients were thresholded with a soft-thresholding method (see Mallat, 1999, for details) in order to reduce signal noise by applying low smoothing. We decided to keep the wavelet coefficients of every resolution level, from which the original spectrum can be rebuilt. In the data set described in this paper, the number q of wavelet coefficients was equal to 367. Another basis, the Daubechies basis made of smooth trimodal elementary functions, was also used, and gave $q = 388$ wavelet coefficients. A more detailed description of the wavelet decomposition can be found at the Online Supplemental Data.

2.4 Selection of variables for prediction

Many prediction methods are described in the literature. Among the most well-known, the Partial Least Square (PLS, Wold, 1966) and Random Forest (Breiman, 2001) methods use all variables, whereas the Lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005) or sparse PLS (Lê Cao et al., 2008) methods incorporate a feature selection step leading to a reduced number of explanatory variables in the model. Some of these methods (Giraud et

al., 2010, preprint, R package available at
http://w3.jouy.inra.fr/unites/miaj/public/perso/SylvieHuet_en.html) were performed on our data set, and gave similar results in terms of predictive power (not shown).
In the case of high dimensionality of the explanatory variables, a feature selection approach is useful for highlighting a limited number of variables of high predictive importance. In general, retaining in the prediction model only a set of useful variables avoids overfitting, and ensures a smaller prediction error. Any variable selection method could have been used here, either on the raw metabolomic data or on the thresholded wavelet coefficients, in order to select the relevant set of parameters. In both cases, this represents a classical problem for variable selection in a linear model. We decided to present here only the most widely used method: the Lasso technique. Introduced by Tibshirani (1996), the Lasso method is a penalized least squares approach used to solve ill-posed or badly-conditioned linear regressions. The main interest of this approach comes from the fact that the solution leads to a restricted number of non-zero coefficients, this number depending on the value of the regularization parameter.
Identifying the points (buckets) of the metabolomic profile that contribute the most to phenotype prediction can then lead to a biological interpretation step. Indeed, some "peaks" (not all) in the profile have already been identified by biochemists to correspond to specific metabolites (one or more metabolites per peak). In the case of data preprocessing however, a single wavelet coefficient can correspond to a large interval in the metabolomic profile, making further interpretation more delicate. Therefore only lists of biomarkers obtained from raw data are presented in the following sections.

2.5 Estimation of predictive power

The Lasso technique was applied on 3 versions of the data collected for the 27 phenotypes described in the Data subsection: the raw data, the thresholded wavelet coefficients obtained with the Haar basis and with the Daubechies basis.
The parameters of each model (see Models below) were estimated first on a subset of the data (learning set with 400 observations), then performances were calculated on the remaining data set (test set with 106 observations). The regularization parameter was tuned by cross validation on the learning set.
The global procedure (estimation of the set of relevant parameters on the learning set and estimation of performances on the test set) was repeated 100 times on several random splits of the whole data set. These random splits took into account the experimental setting of Table 1. This led to a collection of performance values that could be displayed in a boxplot in order to evaluate the level of accuracy of each method as well as its variability.
Performances were evaluated using the mean squared errors of prediction (MSEP) standardized by the variance of the observations, averaged on the 100 test sets. Note that the

MSEP is not upper-bounded, so it can go to infinity for very low predictive powers. However, the lower is the MSEP, the better is the predictive power. A Kolmogorov-Smirnov test of distribution equality was computed for the MSEP on the 100 replicates to test whether two methods were comparable. Paired t-tests were used to test the superiority of one method on another in terms of MSEP.
To achieve a more detailed comparison between the results of all tested methods, we counted the number of appearances of each selected variable (bucket, Haar coefficient, Daubechies coefficient resp.) over the 100 replications, for each data set (raw data, wavelet coefficients obtained either with Haar basis or with Daubechies basis, resp.).

2.6 Models

We focused on three different problems in this paper: the prediction of a phenotype based on the metabolomic data alone (Model 1), based on breed information and the metabolomic data (Model 2), and finally based on batch and breed information and the metabolomic data (Model 3). We considered a linear relationship between a phenotype and the explanatory variables in all three models described above.
Model 1 had the following explanatory variables: Intercept (always in the model) and the metabolome variables (subject to variable selection: 375 for raw data, 367 or 388 for wavelet coefficients with Haar or Daubechies, respectively). Model 2 included a breed effect (always in the model), and the following effects that were subject to variable selection: metabolome variables and breed $\times$ metabolome interactions. Finally, Model 3 included breed and batch effects (both always in the model), as well as metabolome variables, breed $\times$ metabolome and batch $\times$ metabolome interactions (subject to variable selection).

2.7 Canonical analyses

Complementary statistical analyses were performed by regularized canonical analysis using the R package mixOmics (Lê Cao et al., 2009). Two data sets consisting in phenotypic variables and metabolomic variables were represented to evidence the maximal correlations between variables, both within and between the two data sets.

3 Results

3.1 Comparison of models

For all phenotypes, the models based on a wavelet preprocessing step were in general slightly better or at least equal in terms of prediction error, than the one based on the direct use of raw metabolomic data (Figures S5-7 on Supplemental Material). The efficiency of the preprocessing step was most obvious when only metabolomic information was considered

in the model (Model 1). This is well exemplified in the 3 data versions of DFI, both in terms of MSEP and number of selected coefficients (Figure 2), using only the metabolomic information as explanatory variables (Model 1). Indeed, MSEP values were observed to decrease, as was the median number (and strikingly the range) of selected coefficients of the Lasso regression, when wavelet preprocessing of data using the Daubechies basis, but not the Haar basis, was applied. This was corroborated by the comparison of preprocessing methods given by the Kolmogorov-Smirnov test for the MSEP (p-values for raw data vs Haar = 0.58, raw vs Daubechies = 1.3 10^{-5}, Haar vs Daubechies = 1.6 10^{-2}). Thus, transformation of the signal with wavelets implied significant differences in the prediction errors for DFI. Moreover, the results also showed that a phenotype of interest such as DFI could be well predicted with no call for any additional information on the individuals.
When looking into which pre-processing methods gave the best MSEP on average over all phenotypes, no clear conclusion appeared for Model 1 (Figures S11-12), but Daubechies was overall to be preferred to Haar for Model 2 (Figures 4 or S6, S11-12) and Model 3 (Figures S7, S11-12). Moreover, the wavelet transform with the Haar basis gave numerous extreme results in terms of MSEP. This was more detectable in Model 2 than in Model 1 (Figures S5-6). Finally, the p-value of the two-sided Kolmogorov-Smirnov test was equal to 4.10^{-5} for DFI, meaning that there can be a significant difference due to pre-processing in the prediction results for some phenotypes.

3.2 Prediction of phenotypes related to animal breeding and carcass characteristics using metabolomic data

The variation of the prediction levels among all phenotypes was very similar whatever the statistical method used. We present here the results obtained using (i) the best wavelet transform (with Daubechies basis), and (ii) the simplest approach, namely the Lasso method applied to the raw data set (Table S2 and Figure 4), hence retaining the possibility for a more direct biological interpretation of the results than when a wavelet transform pre-processing step is applied (see below). The mean prediction errors (expressed in phenotypic variance units) varied from 0.3 to more than 1. The worst predictions (highest values of MSEP) are obtained for weights measured near slaughter time (i.e. LWETP, CWwtH, HCW, CW, and LWS) and for some phenotypes related to post-mortem meat processing (i.e. pH24 and L*). For LMP, which was the best predicted phenotype with a MSEP value of approx.. 0.3, the squared correlation ($R2$) between observed values and fitted values obtained on the training sample set was equal to 0.82. A $R2$ value between observed and predicted values of 0.69 was observed for the test sample set, showing a good adequacy between observations and adjustments from the model (Figure 3). Use of more complex models was useful to obtain higher prediction scores for some traits as described hereafter.

Reinforced phenotypic prediction using both metabolomic and breed information (Model 2) The phenotypes considered here could be sorted into 4 classes depending on their level of predictability as shown in Figure 4, ranging from the best (class C1 with a MSEP lower than 0.2) to the lowest (class C4 with a relative error rate higher than 0.70). All phenotypes belonging to the classes C1 and C2 were better predicted when the breed was considered in the model (Table S2, Figures 4, S8-10). Prediction using breed, batch and metabolomics information (Model 3) The batch variable does not appear to be a key parameter in the prediction of phenotypes (Table S2, Figures 4, S8-10). Indeed, MSEP values were almost always slightly higher when the batch was taken into account (except shW and DP, for phenotypes of classes C1 and C2).

3.3 Selected variables

As shown in Figure 2B for the DFI phenotype, the number of selected coefficients was always smaller for preprocessed data using a wavelet transform than for raw data. Such transformed data sets gave more parsimonious models with lower numbers of explanatory variables.
Concerning Model 2, it should be recalled that the breed effect did not undergo feature selection; in this setting, the minimum number of selected variables is 3. A non-empty set of metabolites is still of predictive importance, additionally to the breed effect. For Model 3, the breed and the batch did not undergo selection; in this setting, the minimum number of selected variables is 11. The number of selected variables (metabolites and interactions, i.e. breed $\times$ metabolome and batch $\times$ metabolome) is lower when the batch variable is not considered. It is to be noted that no interaction term between metabolites/wavelet coefficients and breed (or batch) was selected in Model 2 (or in Model 3).

A few of the explanatory variables obtained for the prediction of the LMP phenotype (Table 3) were the same when using raw data (Model 1) as when using Models 2 or 3. However, their number was significantly reduced when the breed factor is taken into account in Models 2 and 3 compared to Model 1. When using the bootstrap process, some variables were either mostly positively linked (PL) (i.e.δ 4.05 ppm, 2.43 ppm, 2.15 ppm, 1.33 ppm and 1.45 ppm), negatively linked (NL) (i.e.δ 3.93 ppm, 3.20 ppm, 7.67 ppm, 2.51 ppm and 0.99 ppm) or both positively and negatively linked (δ 1.03 ppm, 2.25 ppm, 1.47 ppm) to LMP (not shown). Only variables that are steadily linked, either positively or negatively, such as creatinine (δ 4.05 ppm, PL), creatine (δ 3.93 ppm, NL), choline / phosphocholine / glycerophosphocholine (δ 3.20 ppm, NL), glutamine (δ 2.43 and 2.15 ppm, PL), lactate (δ 1.33 ppm, PL), alanine (δ 1.45 ppm, PL), and isoleucine (δ 0.99 ppm, NL) can be considered for the elaboration of the functional hypotheses that could explain

how the LMP phenotype can be predicted from these serum biomarkers. Interestingly, as displayed in Figure 5A, canonical analysis performed on all the variables present in the two data sets (i.e. 1H NMR and phenotype ones) demonstrated that the phenotypic variables belonging to the classes 1 and 2 were also those that were steadily selected in Models 1, 2 and 3. So, the positive correlation underlined by the Lasso-based regression between LMP and creatinine (δ 4.05 ppm) or glutamine (δ 2.43 ppm) is again well evidenced, as is the negative link between LMP and creatine detected atδ 3.93, 3.92 and 3.03 ppm (Figure 5B). This significant correlation between LMP and creatine is also well evidenced for class 2 phenotypes such as ComLMP, DP, shW, hamW, beW and DFI (Figure 5B). Citrate would be also found as NL regressor of LMP when considering the chemical shift atδ 2.51 ppm in Model 1, but would be found as PL regressor of LMP if we consider the variable atδ 2.54 ppm. 2D 1H-1H COSY and 1H-13C HSQC NMR spectra showed that signals at 2.51 ppm and 2.54 ppm are belonging to citrate. Indeed, HSQC NMR spectra showed correlation between 13C chemical shift at 48.6 ppm and 1H chemical shift at 2.51 and 2.54 ppm. Chemical shift atδ 2.51 and 2.54 ppm have been assigned to citrate and correspond to a doublet even the chemical signal recorded at δ 2.54 ppm, that may contain also a low intensity signal attributable to β-alanine(correlation between the signals at 3.17 and 2.54 ppm in the COSY spectrum) and an unknown compound (correlation between the signals at 2.39 and 2.54 ppm in the COSY spectrum). Quantitative information measured at these two chemical shifts are correlated ($\rho = 0.35$) and would be in favor of an assignment to citrate, even though the correlations with LMP are of different signs, but based on different models involving very different numbers of regressors (Table 2).

3.4 Reasoning at constant weight

There was some variability in the development status of the pigs included in the data set, both at the time of blood sampling and at the time of slaughtering. In order to be able to compare samples, the weight of the animal at slaughter time (LWS) was added as covariable in the 3 models described previously. Then the phenotype prediction could be considered as being at constant weight. Focusing on the LMP phenotype, the results obtained with these 3 modified models were similar in nature to those presented previously: the knowledge of the breed improved the prediction of the phenotype and decreased the number of explanatory variables selected. Moreover, the relation between LMP and the few variables referred to above (PL or NL) was preserved. More precisely, the lists of important metabolites were larger and included those already highlighted in the model that did not take into account the animal weight. However, the prediction power was slightly lower when the weight at slaughter time was considered (not shown).

4 Discussion

In this paper, we showed that it is possible to use metabolomic data from a plasma sample to better predict some production phenotypes in the growing pig. Metabolomic data alone are sufficient to predict these phenotypes. Additional information and predictive power are provided by the metabolome when the breed of the animal is known. For data from a test farm, micro-variations in a breeding environment (that are classically summarized in a batch effect) did not disrupt phenotype predictions. Additionally, although this work was centered on prediction accuracy, we supplied supplementary information on a limited number of metabolites that have, as valuable biomarkers, a high predictive power. The biological coherence of the list of biomarkers validated somehow the whole data analysis. In addition, a methodological aspect of the statistical treatment was related to the specificity of 1H NMR metabolomic data: a pre-treatment of the signal based on the use of wavelets.

4.1 Justification of the statistical treatment

Metabolomic profiles are continuous by essence. Discretization is performed routinely (bucket steps). The bucket size was rather large with 0.01 ppm, to avoid a possible misalignment between spectra, due to shifts of signals, a rather rare phenomenon but still occurring. Actually, small shifts at 2-3 regions of the spectrum recorded in plasma samples were locally observed for some samples that were reanalysed by the same spectrometer at 2 different times (not shown). This motivated the choice of a relatively large bucket size (0.01 ppm), even though a consequence is that some buckets could contain more than one compound. All the more as the primary goal of this work was prediction and not biological interpretation.

To recover the continuity of the signal, that is moreover non-regular, we proposed the use of wavelet decomposition, which is one of the most commonly-used signal transformation approaches. The underlying idea is to decompose a complex signal into elementary forms (orthogonal functions, or basis). Unlike Fourier transformation, the wavelet approach is particularly suited for uneven and chaotic signals, making it a method of choice for NMR profiles and it has already been applied in such a context by Davies et al. (2007) and Xia et al. (2007). An improvement due to the used of wavelet transformation was observed on our data, but in a limited manner. Depending on the tissue (blood, urine, other), the stability of the baseline on the spectra, the wavelet approach could lead to a dramatic improvement of the signal (Martin, Besse, Déjean, personal communication; Villa-Vialaneix, Paris et al, in prep.): approximations of the signal at the lowest levels (see Supplemental Material) correct rough fluctuations of the baseline. Results depended on the chosen wavelet basis in this study, but only slightly. When the signal is continuous, Daubechies wavelets are usually a better choice than Haar ones (step functions). The dependency on the basis is

generally observed (e.g. Luisier et al. (2005) for image denoising, Mahmoud et al. (2007) for audio data, etc...).

4.2 Predictive power: valuable aspects for all phenotypes

An important methodological question arose prior to the global prediction analysis concerning the choice of preprocessing the 1H-NMR metabolomic spectra. When considering metabolomic data only as predictive variables of highly functionally integrated phenotypic variables, as shown here, the wavelet transformation of original data led to best performances.

Adding information concerning the breed led to lower errors of prediction, while adding batch information did not really improve the prediction results. More, the batch even seemed to constitute a noisy endogenous variable as the predictive power in Model 3 was slightly lower than in Model 2. Interestingly, in the breeding conditions encountered here, this meant that we could put aside the possible micro-environmental effect (that may vary from batch to batch) for a phenotype prediction objective. The environmental effect on the phenotype, particularly diet variation, is probably captured by the metabolomic information (Yde et al., 2010). Thus, given the fact that data are obtained in a control farm that ensures standardized breeding conditions, some phenotypes of interest such as LMP can be well predicted without having to characterize more precisely the micro-environment of a given batch of growing individuals. The same phenomenon seems to be encountered for the slight variations of animal weight or age that were observed in the data set: the metabolome carries some information pertaining to developmental differences, so that the prediction of some phenotypes such as LMP is better without the weight information than with it.

Yet, this conclusion is based on a large data set issued from 3 breeds. Indeed, when similar analysis was undertaken within a given breed, predictions of phenotypes were disastrous (not shown). This can be explained by the lower number of observations and by a lower variability of the within-breed phenotype as can be seen in Figure 3 for instance.

4.3 Prediction power among phenotypes and practical implications

The prediction accuracy is very dependent on the phenotype being studied, and surprisingly even within a group of related phenotypes. Canonical analysis confirmed the Lasso-based predictions and the same 4 classes of prediction of the different phenotypes were identified (Figure 5). Two groups of phenotypes were badly predicted (class 4 of prediction). They correspond to:

- Some weights (LWETP, LWS, CWwtH), the values of which depend directly on the

decision to send animals to the slaughterhouse or not. Therefore, these phenotypes can be considered as negative controls, because they should be badly predicted by essence, and not worth predicting.

- Meat quality measurements (pH24, L*, a*, b*, WHC, MQI). The bad predictions obtained for these phenotypes can be easily explained by the fact that meat quality is highly influenced by pre-slaughter conditions, whereas the blood sample was collected at the test farm during the growing period between 60 and 70 kg BW. Indeed, the pH is known to be very sensitive to the duration of fasting, transportation, etc. Moreover, evidence of stress conditions has been observed on NMR metabolomics in pigs (Bertram et al., 2010) near slaughter, or in sheep (Li et al., 2011). Meat quality, even though it does not represent a direct objective for the selection because it is difficult to measure, could be potentially considered as a prime objective if good predictions were available. Metabolomic data from a single blood sample, taken approximately 3 weeks prior to slaughter, are clearly not sufficient for such an ambitious task for this complex trait.

Backfat measurements (BFsh, BFlr, BFhj and their average mBF) all showed a medium level of predictability (class 3 of prediction), potentially linked to the dynamics of fat deposition during growth, which essentially occurs after 70 kg BW. However, the metabolome-based prediction of these phenotypes is not crucial since they are easily measured on the living animal. The carcass length (Length) displayed also a limited prediction level, but is of no economic interest to date. In the last 3 groups of phenotypes, one phenotype within each group was accurately predicted, while the others were not:

- Concerning traits recorded during growth (ADG, FCR, DFI), we observed that DFI was better predicted than ADG and FCR separately. Individual measurements of DFI require specific and expensive equipment, and are hence rarely performed. However, it represents a very important criterion from an economic perspective, and presents a medium to good level of prediction here.

- As regards to carcass efficiency, DP was actually quite well predicted (class 2 of prediction), even though individual weights (CW and LWS) were not.

- The lean meat content estimated from cut weights (LMP) displayed the highest prediction accuracy (class 1 of prediction). The prediction of separate pieces weights varied from bad to good, but was always worse than LMP. Lean meat content is a crucial parameter for the breeders since it directly influences the payment of carcasses. Two measurements were available and ComLMP and LMP are highly correlated (Figure 5). The latter measurement is time-consuming and requires half of a carcass for the cutting of the various pieces. The LMP impacts the income of the breeder

and the slaughterhouse, and displayed the highest predictability level among the phenotypes considered here, as well as among those included in the current selection objective (i.e. MQI, ADG, FCR and LMP).

4.4 A possible biological interpretation of the good prediction performance of LMP

The purpose of this work was not to dissect the metabolism mechanisms linked to the measured traits, but to quantify the power of prediction of NMR metabolomic spectra for production and quality traits. Discussing biological aspects of the most predictive metabolites can be proposed, but only to check biological coherence of the whole statistical process. Because of a risk of over-interpretation, we chose to limit the discussion on that point. The results obtained above can thus be validated considering the coherent biological significance of the metabolites selected to predict LMP. Indeed, a connection between the phenotype LMP and some metabolites found in plasma has been highlighted. It involves (i) three amino acids: valine, alanine and glutamine, (ii) an energetic intermediate of the Krebs cycle, citrate, (iii) an end metabolite of amino acids, creatinine, and its precursor creatine, and (iv) choline, a quaternary ammonium derivative, involved in the biosynthesis of the choline-containing phospholipids, acetylcholine and betaine.

In Model 1, the lean meat percentage (LMP) measured at slaughter is positively linked to circulating creatinine and negatively linked to creatine measured between 60 and 70 kg BW. Creatinine is directly linked to the muscular mass and as such is correlated to the total amino acid catabolism in muscle, which may depend on gender and hormonally-based anabolic treatment (Dumas et al., 2005). Interestingly, when no qualitative covariate such as "breed" (Model 2) or "batch" (Model 3) is used in the prediction model, creatine is found in plasma as an independent variable negatively linked to LMP. This may imply that the energetic requirements needed to sustain muscular metabolism are adjusted in a coordinated manner according to the relative potential to increase the muscle mass, and result in different circulating concentrations of creatine. When breed or batch covariates are introduced in the models, creatine is not found as a main independent variable. Probably, creatine as precursor of phosphocreatine – this phosphagen represents the greater part of the total P-bonded energy in muscle instantaneously available to regenerate ATP (Hochachka, 1994; Brosnan and Brosnan, 2007) – is metabolized at different levels in the different breeds, as it seems to be linked to a final LMP phenotype which is strikingly differentiated between breeds and probably between genders. Glutamine, detected atδ 2.43 ppm, and lactate, detected atδ 1.33 ppm, also displayed a differential pattern of energy supply to muscle which was positively correlated to LMP between breeds (and genders). Glutamine, as functional amino acid, is involved in multiple metabolic pathways and regulates gene expression and signal transduction pathways (Wu, 2010; Wu et al., 2011). Among

its different physiological functions, it is an important energy substrate, more particularly for rapidly dividing cells such as enterocytes. Intra-breed (and -gender) variations in LMP are also positively correlated to citrate. As for phosphagen P-creatine, a higher potential in muscle accretion seems to be coordinately sustained by systemic bioenergetic adaptation observed at the level of the citric acid cycle and lactate metabolism. Unfortunately, complementary observations are lacking so it is difficult to provide, at this stage, sound physiological interpretation concerning the relative involvement of factors related either to the genetic background or to a gender-adjusted physiology of such energetic homeostatic adjustments. Indeed, there are here two confounded factors leading to LW or LR castrates on one side and PI females on the other.
As the data (raw, Haar transformed or Daubechies transformed) may have some influence on the selected metabolites, we displayed on the mean spectrum the regions corresponding to the selected variables (Figure S3), on the particular case of Model 2 for the LMP phenotype as a matter of example. These results showed that the use of raw data is the best approach if one is interested in a biological interpretation, while the pre-processing using the Daubechies basis is overall the best approach in the case of prediction (even though its effect is not tremendous on our data set). The pre-processing with the Haar basis appeared as a trade-off between the 2 goals: biological interpretation and phenotype prediction.
The 3 approaches all pointed out the fine region of the spectrum corresponding to the creatinine (4.05 ppm). The selected points of the raw data (Figure S3a) were included in the larger regions pointed out by Daubechies (Figure S3c), which displayed too large regions to be interpretable.
The purpose of this paper was to predict a phenotype with NMR metabolomic profiles. This is different from an analysis aiming at dissecting the phenotype and discovering metabolites underlying the trait. We only proposed a discussion on the selected metabolites (those with the highest predictive value) for the sake of biological coherence. In this context, it is not a problem that the same metabolites are selected for two highly correlated phenotypes. This could be due (or not) to a common set of metabolism mechanisms.

Metabolomic profiles are now relatively cheap. One may use them in practice to obtain targeted metabolic information for identified biomarkers, or to predict phenotypes of economic interest. Several samples could be considered during the animal's life, depending on the phenotypes desired (i.e. linked to growth during the breeding period, or linked to meat quality near slaughter time). Generally speaking, metabolomic-based prediction of production phenotypes would be of practical interest in animal selection, especially when phenotypes cannot be measured directly on selection candidates, since the measurements require slaughter (carcass efficiency traits, meat quality traits), or are too expensive (feed efficiency). The current solution is to measure these traits on relatives of selection candidates, and this information is used to predict the genetic value of the candidates. However,

phenotypic measurements performed on the animal itself rather than on its relatives, would provide more accurate predictions of the genetic value. If individual meat quality traits could be predicted by accurate indirect measures (based on metabolome profiles), selection would be more efficient than when based on the performances of relatives (which is, moreover, more expensive). The first results obtained in this study need further validation before any practical use in selection schemes.

In conclusion, metabolomic data can be used to predict a phenotype without any further knowledge of the individual. Nevertheless, this prediction ability is again improved when the breed information is available as additional data. For prediction purposes in general, a well-adapted method of reducing noise in data coupled with a sparse prediction approach is to be recommended. This is the first time to our knowledge that breeding and production traits on the growing pig have been predicted on the basis of a single blood sample collected on the living animal during its breeding period. The prediction accuracies varied considerably among the traits, and some of them showed indeed an accurate prediction. We are enthusiastic on the finding that some main economically important traits can be predicted from a simple NMR metabolomic profile achieved on blood.

LITERATURE CITED

Anonymous. 1990. La nouvelle découpe normalisée. Techni-Porc 13(5): 44-45.

Bertram, H.C., N. Oksbjerg, and J.F. Young. 2010. NMR-based metabonomics reveals relationship between pre-slaughter exercise stress, the plasma metabolite profile at time of slaughter, and water-holding capacity in pigs. Meat Science 84: 108–113.

Breiman, L. 2001. Random forests, Machine Learning 45: 5-32.

Brosnan, J.T, and M.E. Brosnan. 2007. Creatine: endogenous metabolite, dietary, and therapeutic supplement. Annu. Rev. Nutr. 27:241-261.

D'Alessandro, A., C. Marrocco, V. Zolla, M. D'Andrea, and L. Zolla. 2011. Meat quality of the longissimus lumborum muscle of Casertana and Large White pigs: metabolomics and proteomics intertwined. J. Proteomics 75:610-627.

Daumas, G., D. Causeur, T. Dhorne, and E. Schollhammer. 1998. Les méthodes de classement des carcasses de porc autorisées en France en 1997. Journées de la Recherche Porcine en France 30:1-6.

Davis, R., A. Charlton, J. Godward, S. Jones, M. Harrison, and J.C.Wilson. 2007. Adaptive binning: an improved binning method for metabolomics data using the undecimated wavelet transform, Chemometrics and Intelligent Laboratory Systems 85:144-154.

Dumas, M.E., C. Canlet, J. Vercauteren, F. André, and A. Paris. 2005. Homeostatic signature of anabolic steroids in cattle using 1H-13C HMBC NMR metabonomics. J. Proteome Res. 4:1493-1502.

He, Q., P. Ren, X. Kong, Y. Wu, G. Wu, P. Li, F. Hao, H. Tang, F. Blachier, and Y. Yin. 2012. Comparison of serum metabolite compositions between obese and lean growing pigs using an NMR-based metabonomic approach. J. Nutr. Biochem. 23:133-139.

Hochachka, P.W. 1994. Muscles as molecular machines. CRC Press, Boca Raton.

Institut Technique du Porc. 1993. Le nouvel IQV. Internal document, 2p.

Labroue, F., R. M.C. Guéblez, Meunier-Salaün, and P. Sellier. 1993. Alimentation électronique dans les stations publique de contrôle des performances : paramètres descriptifs du comportement alimentaire. Journées de la Recherche Porcine en France 25:69-76.

Lê Cao, K.-A., I. González, and S. Déjean. 2009. IntegrOmics: an R package to unravel relationships between two omics data sets. Bioinformatics, 25:2855-2856.

Lê Cao, K.A., D. Rossouw, C. Robert-Granié, and P.Besse. 2008. A sparse PLS for variable selection when integrating Omics data. Stat. Appl. Genet. Mol. Biol.7:Article 35.

Li, J., G. Wijffels, Y. Yu, L.K. Nielsen, D.O. Niemeyer, A.D. Fisher, D.M. Ferguson, and H.J. Schirra. 2011. Altered Fatty Acid Metabolism in Long Duration Road Transport: An NMR-based Metabonomics Study in Sheep. J. Proteome Res. 10:1073–1087.

Luisier, F., T. Blu, B. Forster, and M. Unser. 2005. Which wavelet bases are the best for image denoising?, Proceedings of the SPIE Conference on Mathematical Imaging: Wavelet XI, San Diego CA, USA, SPIE, July 31-August 3, 2005.

Mahmoud, M.I., I. M.M. Dessouky, S. Deyab, and F.H. Elfouly. 2007. Comparison between Haar and Daubechies Wavelet Transformions on FPGA Technology. World Academy of Science, Engineering and Technology 26.

Mallat, S. 1999. A wavelet tour of signal processing. Academic Press, San Diego, USA.

Métayer, A. and G. Daumas. 1998. Estimation, par découpe, de la teneur en viande maigre des carcasses de porcs. Journées Rech. Porcine en France, 30:7-11.

Rochfort, S. 2005. Metabolomics Reviewed: A New "Omics" Platform Technology for Systems Biology and Implications for Natural Products Research. J. Nat. Prod. 68:1813–1820.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. J. Royal Statist. Soc. B 58:267-288. Wold, H. 1966. Estimation of principal components and related models by iterative least squares. Multivariate analysis. New York

Wu, G., F.W. Bazer, G.A. Johnson, D.A. Knabe, R.C. Burghardt, T.E. Spencer, X.L. Li, and J.J. Wang. 2011. Triennial Growth Symposium: important roles for L-glutamine in swine nutrition and production. J. Anim. Sci. 89:2017-2030.

Wu, G. 2010. Functional amino acids in growth, reproduction, and health. Adv Nutr 1:31-37.

Xia, J.M., X.J. Wu, and Y.J. Yuan. 2007. Integration of wavelet transform with PCA and ANN for metabolomics data-mining. Metabolomics 3:531-537.

Yde, C.C., H.C. Bertram, and K.E.B. Knudsen. 2010. NMR-based metabonomics reveals distinct metabolic profiles of plasma from sows after consumption of diets with contrasting dietary fiber levels and composition Original Research Article. Livest. Sci. 133:26-29.

Zhang, A., H. Sun, P. Wang, Y. Han, and X. Wang. 2012. Recent and potential developments of biofluid analyses in metabolomics. J. Proteomics 75:1079-88.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net, J. Royal Statis. Soc. B 67: 301-320.

	Batch							
	1	2	3	4	5	6	7	8
Large White, dam breed	42	45	54	13	16	20	9	0
Landrace, dam breed	22	39	51	0	21	28	27	0
Pietrain, sire breed	0	37	29	5	0	33	0	17

Table 1: Number of pigs in every breed × batch combination.

Model 1		Model 2		Model 3	
δ (ppm) (n)	Assignment	δ (ppm) (n)	Assignment	δ (ppm) (n)	Assignment
4.05 (100) PL	creatinine	4.05 (100) PL	creatinine	4.05 (100) PL	creatinine
3.93 (100) NL	creatine	1.04 (92) NL	valine	2.25 (97) NL	valine
2.43 (100) PL	glutamine	2.54 (88) PL	citrate, β-alanine, unknown	1.04 (84) NL	valine
1.33 (100) PL	lactate	2.40 (78) PL	glutamine	2.54 (83) PL	citrate, β-alanine, unknown
3.20 (97) NL	choline, P-choline, glycerol-P-choline	2.25 (78) NL	valine		
1.45 (89) PL	alanine				
2.15 (82) PL	glutamine				
7.67 (80) NL	unknown				
2.51 (74) NL	citrate				
0.99 (74) NL	isoleucine				

Table 2: Variables selection for Lean Meat Percentage using the raw data for the three models: metabolomic data alone (Model 1), metabolomic + breed (Model 2) and metabolomic + breed + batch (Model 3). Chemical shifts (δ) in ppm and putative assignments are given. The appearance of the variable over the 100 replications is given between parentheses, threshold at 70. Metabolites that are positively (resp. negatively) linked with LMP are denoted by PL (resp. NL).

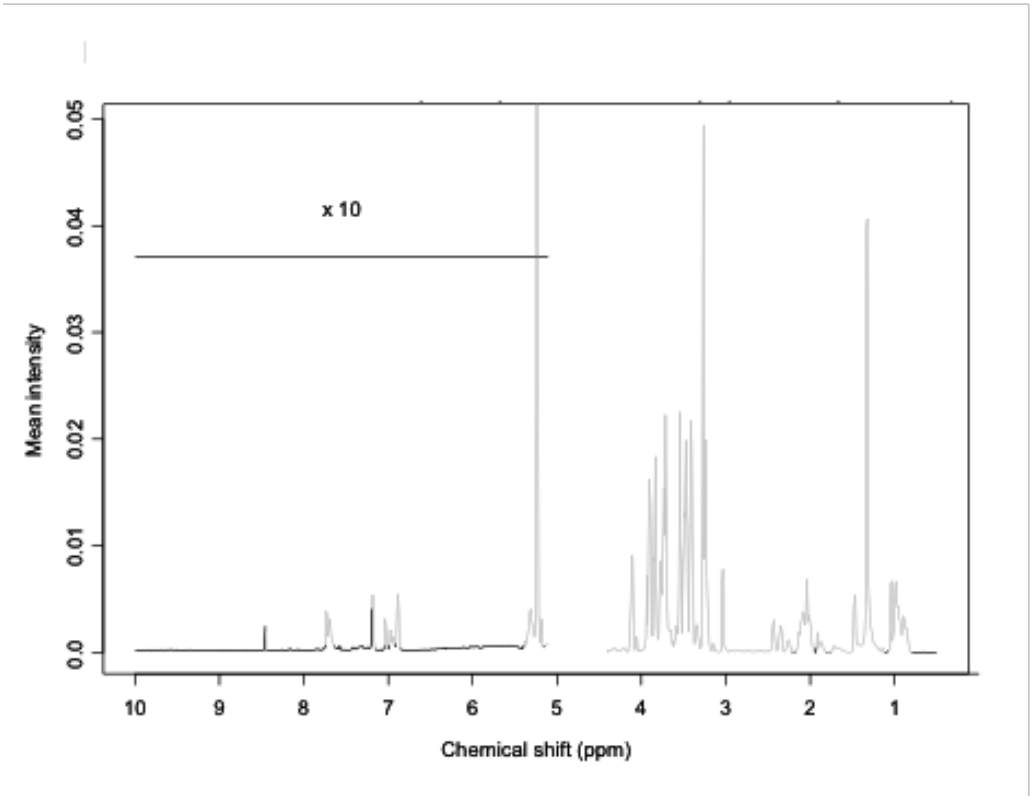


Figure 1: 1H NMR spectrum acquired on plasma collected on one growing pig weighing 60 kg. Informative variables preselected by a multidimensional scaling procedure performed on the transposed matrix of metabolomic data transformed into 0.01-ppm buckets are colored in grey, when residual information found in baseline is colored in black. A 10-fold magnification of the spectrum in the aromatic region above 5.15 ppm is applied.

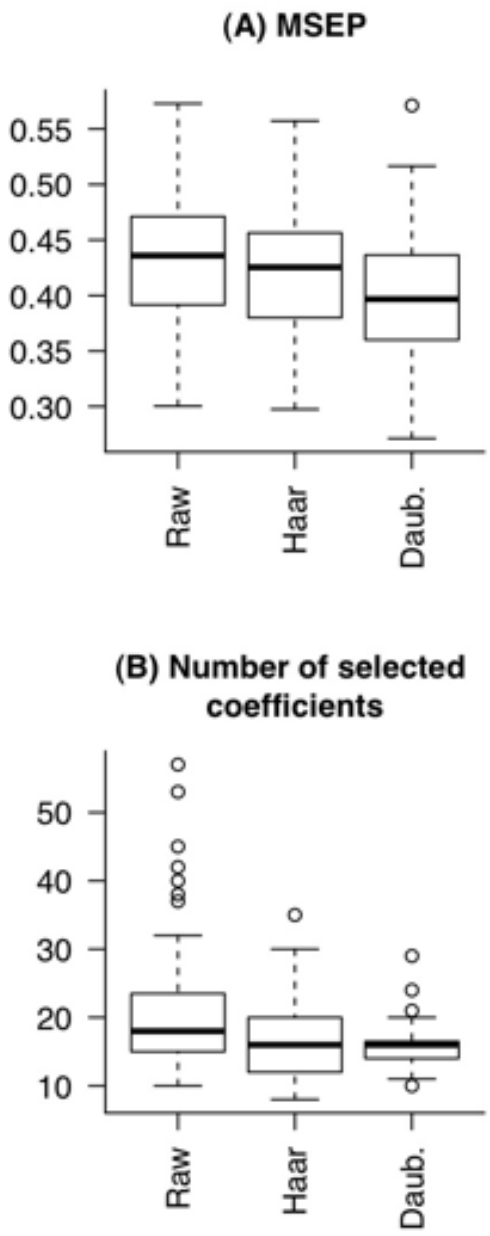


Figure 2: Prediction of Daily Feed Intake. Boxplot of the preprocessing methods considered over 100 resampling replicates, in the model with metabolomic data only, on raw data (Raw), preprocessed data with Haar wavelet transformation (Haar) and Daubechies wavelet (Daub.). (A) Mean Square Error of Prediction (MSEP), (B) Number of selected coefficients.

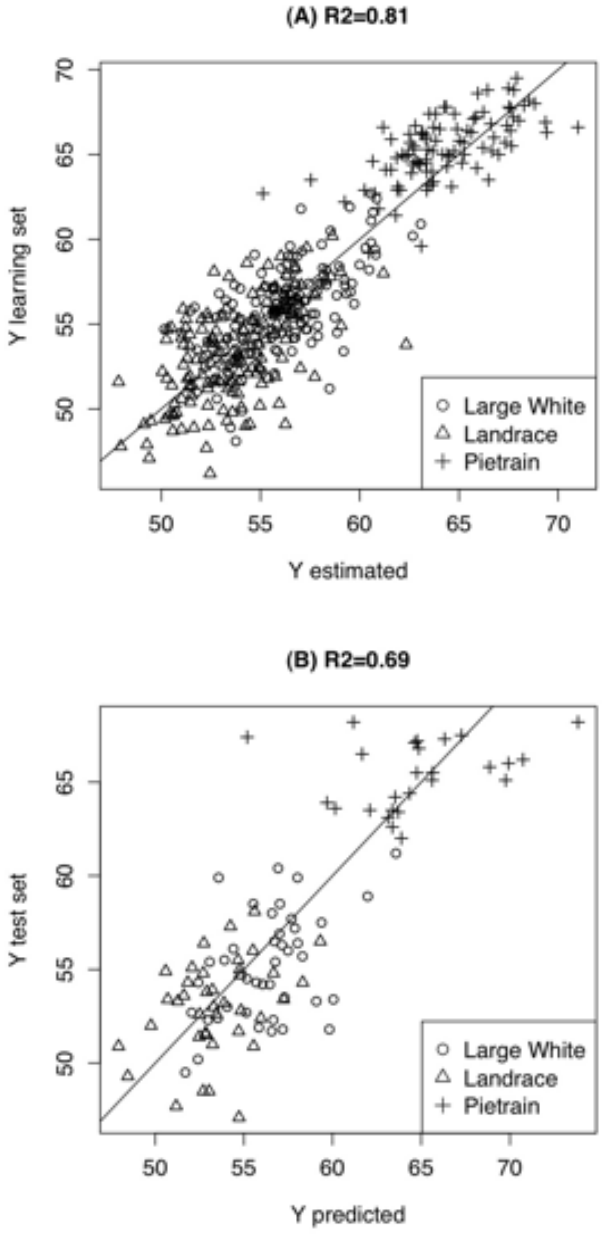
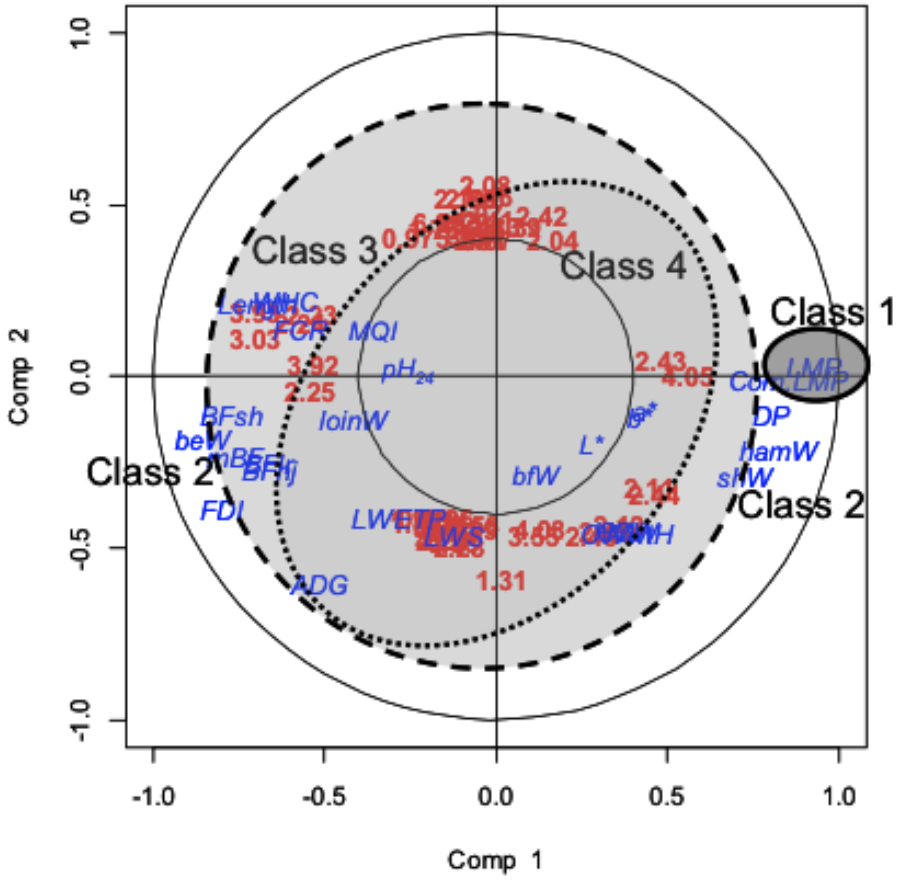


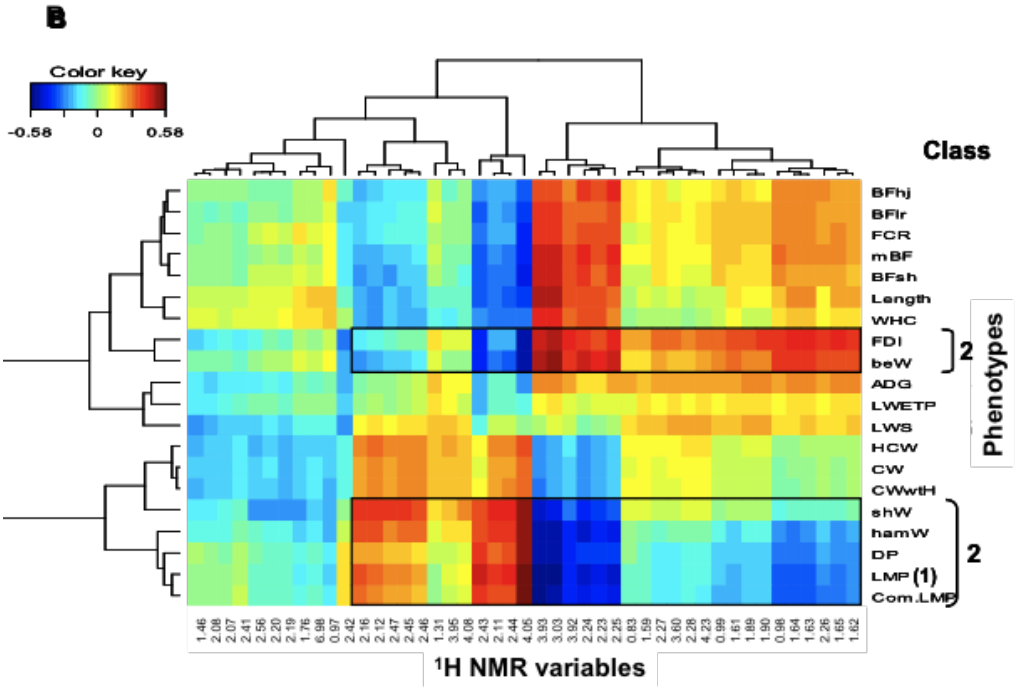
Figure 3: Lean Meat Percentage phenotype. Estimated values on the learning set (A) and predicted values on the test set (B), both against the true values. Predictive model with metabolomic data only (Model 1).

Figure 4: Mean Square Error of Prediction for all the considered phenotypes, on the raw metabolomic data with breed information, expressed in phenotypic variance units. C1, C2, C3 and C4 define 4 classes of prediction accuracies. The 3 pre-processing methods are displayed (raw data, wavelet transformation with Daubechies basis, and with Haar).



(a)

Figure 5: Canonical analysis between the 1H NMR data set (X) and the phenotype data set (Y). a. Projection of variables. 1H NMR variables with correlation less than 0.4 were not plotted. b. Correlation heatmap between variables belonging to the two datasets (X and Y). Classes of variables refer to the prediction levels in Figure 4.


B
Color key
-0.58
0
0.58
Class
BFhj
BFlr
FCR
mBF
BFsh
Length
WHC
FDI
beW
ADG
LWETP
LWS
HCW
CW
CWwtH
shW
hamW
DP
LMP (1)
Com.LMP
2
2
Phenotypes
1.46 2.08 2.07 2.41 2.56 2.20 2.19 1.76 6.98 0.97 2.42 2.16 2.12 2.47 2.45 2.46 1.31 3.95 4.08 2.43 2.11 2.44 4.05 3.93 3.03 3.92 2.24 2.23 2.25 0.83 1.59 2.27 3.60 2.28 4.23 0.99 1.61 1.89 1.90 0.98 1.64 1.63 2.26 1.65 1.62
¹H NMR variables

(b)

Supplemental Material

Wavelet decomposition. There will be several levels of decomposition of an initial spectrum from level N-1 –high resolution- to level 0 –rough tendency-. The number of these levels is N=9 (because the number of buckets p = 375 lies between $2^8 = 256$ and $2^9 = 512$), The initial spectrum f(t) is decomposed as the sum of a detail spectrum D8(t) and an approximation A8(t). Then the approximated spectrum A8 is decomposed into a further detail spectrum D7 and a further approximation A7. Each approximated spectrum is decomposed sequentially as the sum of a detail spectrum and of an approximation spectrum (as a residual), as illustrated in Figure S1 for the Daubechies basis. The detail spectrum of level j is obtained as:

$$D_j(t) = \sum_{k \in \mathbb{Z}} b_{j,k} \psi_{j,k}(t)$$

where each $\psi_{j,k}(t)$ is a translation and a dilatation of the so-called mother wavelet $\psi(t)$ (Haar that is a simple step function, or Daubechies a continuous trimodal function). In practice, the index k is in a finite support. The coefficients b are called the (detailed) coefficients and are equal to $b_{j,k} = \int f(t)\psi_{j,k}(t)$. An empirical estimator of these coefficients is used, from the values of the discretized spectrum at points t_i . Some of the numerous wavelet coefficients are close to 0, so thresholding is made to reduce the number of non-null coefficients.
Similarly, the approximated spectrum of level j is obtained as:

$$A_j(t) = \sum_{k \in \mathbb{Z}} a_{j,k} \phi_{j,k}(t)$$

where each $\phi_{j,k}(t)$ is a translation and a dilatation of the so-called father wavelet $\phi(t)$. The coefficients a are called the approximated coefficients.

The initial signal $f(t)$ can be entirely reconstructed from all detail spectra and the approximation A_0 at the lowest resolution level:

$$f(t) = D_{N-1}(t) + A_{N-1}(t) = D_{N-1}(t) + D_{N-2}(t) + \cdots + D_0(t) + A_0(t)$$

since $A_j(t) = A_{j-1}(t) + D_{j-1}(t)$ for $j \in \{1, \ldots, N-1\}$.
The (detailed) wavelet coefficients b estimated from the data in Figure S1 are plotted in Figure S2 for all resolution levels.

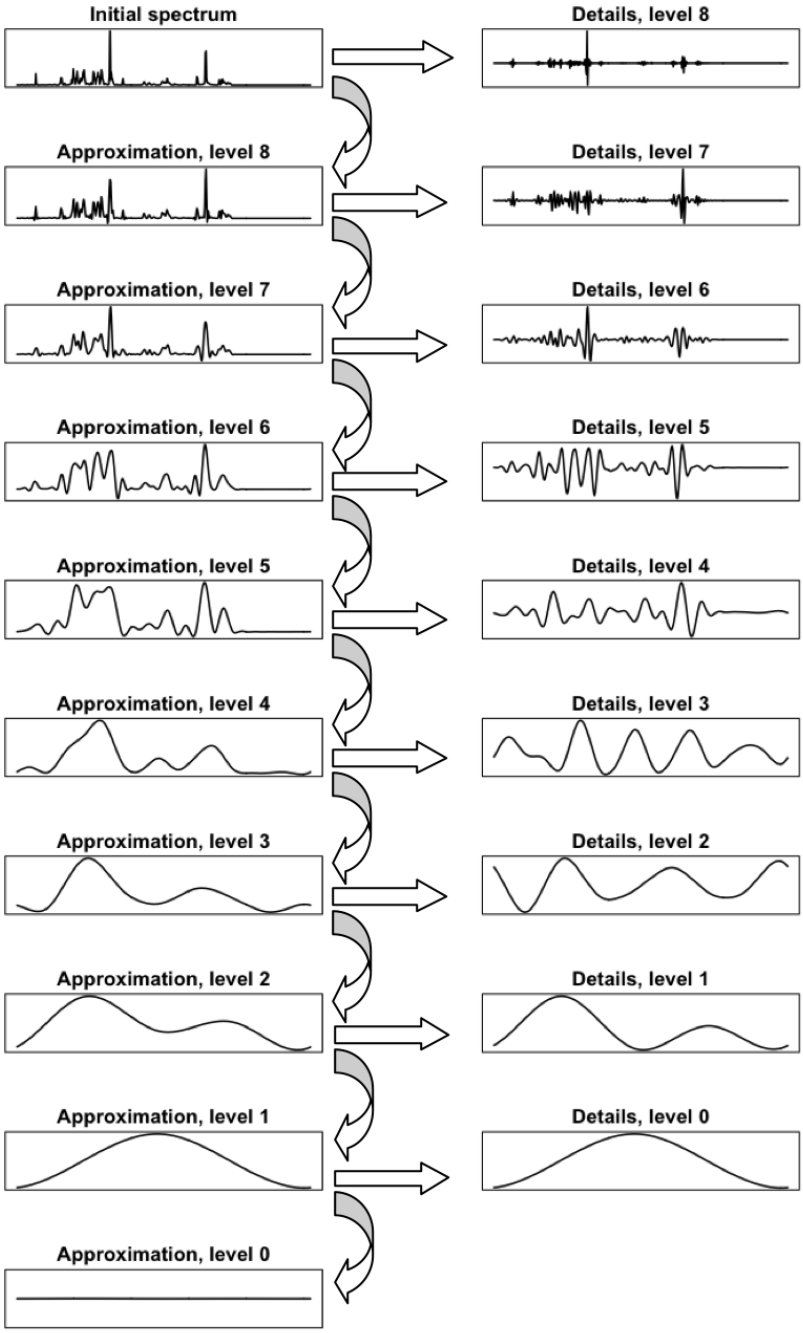


Figure S1. The eight levels of decomposition of the initial spectrum with the Daubechies wavelets.

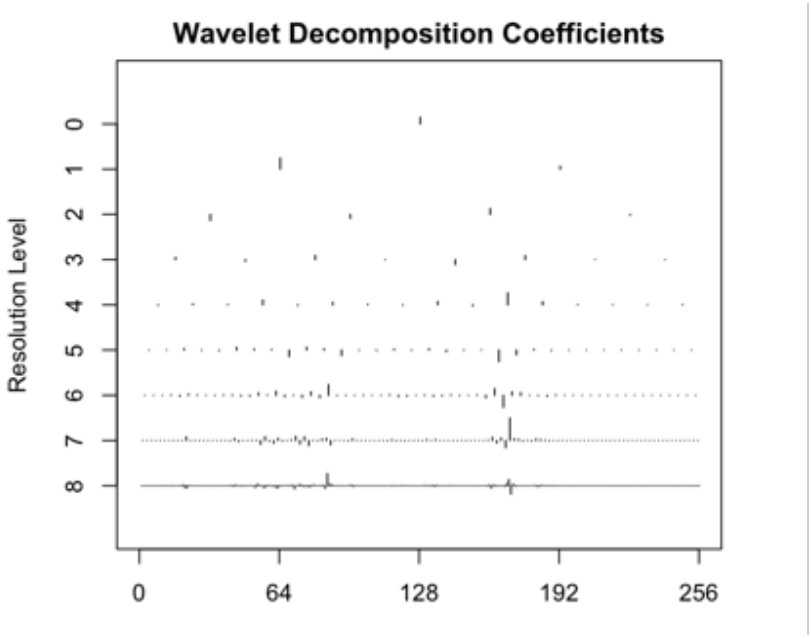


Figure S2. The wavelet coefficients of each level of the Daubechies decomposition.

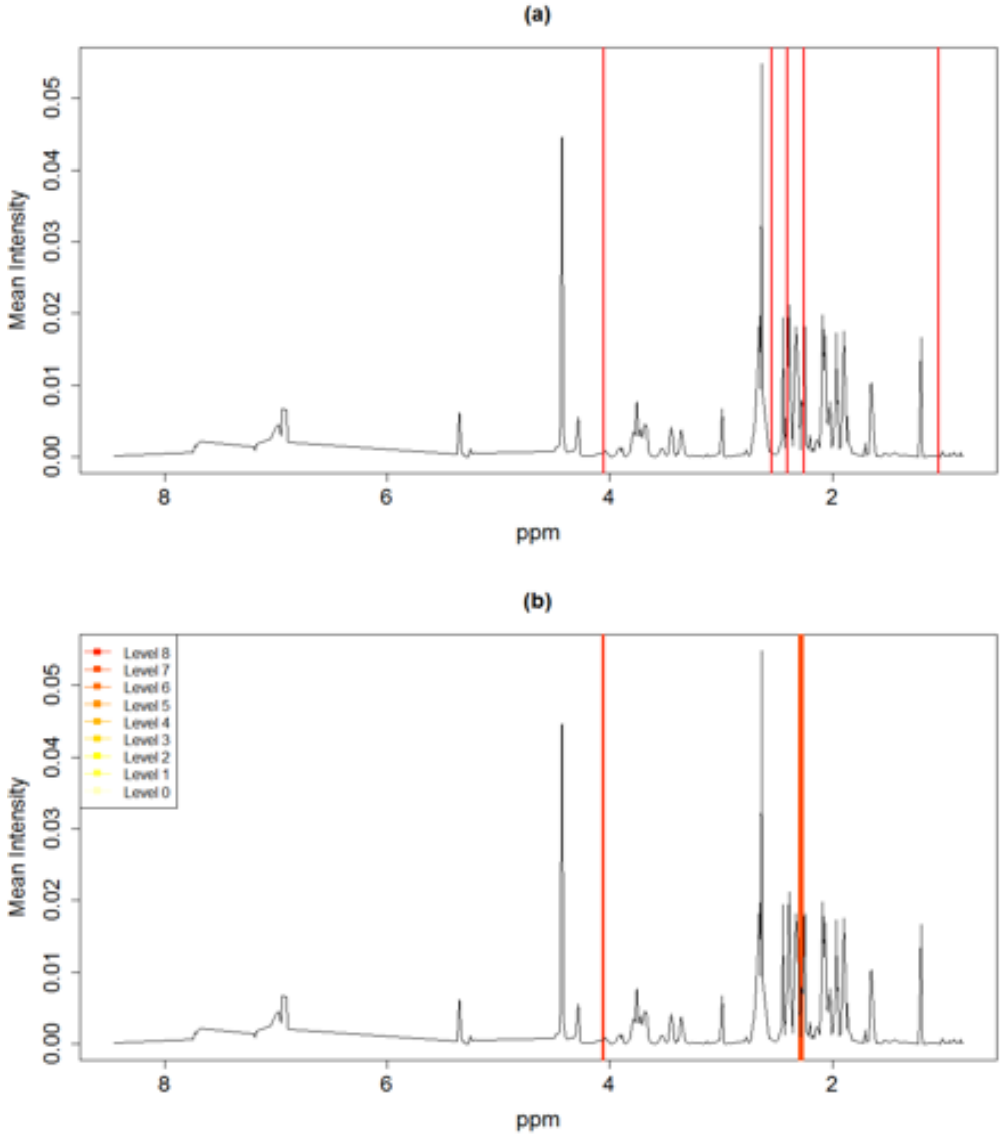
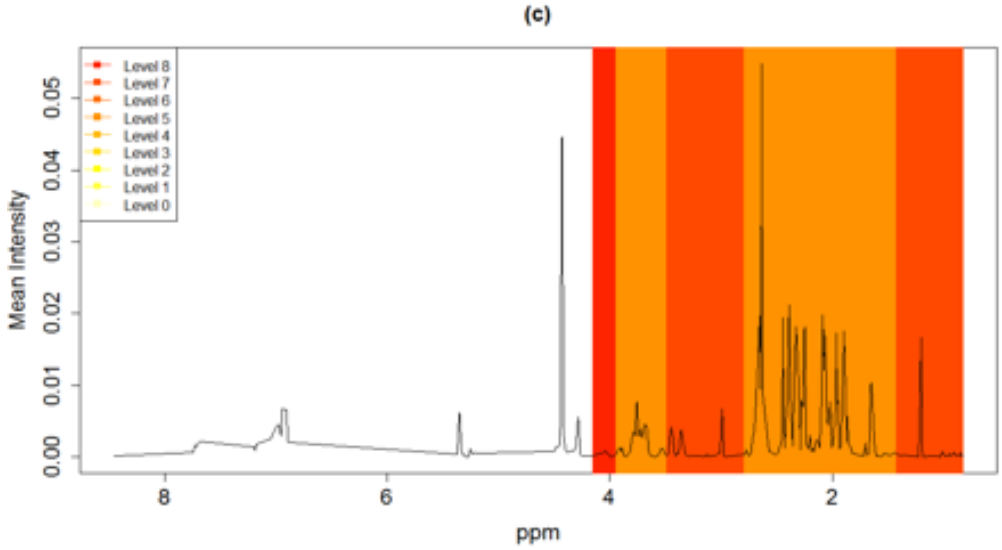


Figure S3. Parts of the metabolomic spectrum that are highlighted by the Lasso method for LMP phenotype in Model 2 (a) for the raw spectrum, (b) for the pre-processed spectrum using the Haar wavelet basis, (c) using the Daubechies basis.


(c)
Level 8
Level 7
Level 6
Level 5
Level 4
Level 3
Level 2
Level 1
Level 0
Mean Intensity
0.00
0.01
0.02
0.03
0.04
0.05
8
6
4
2
ppm

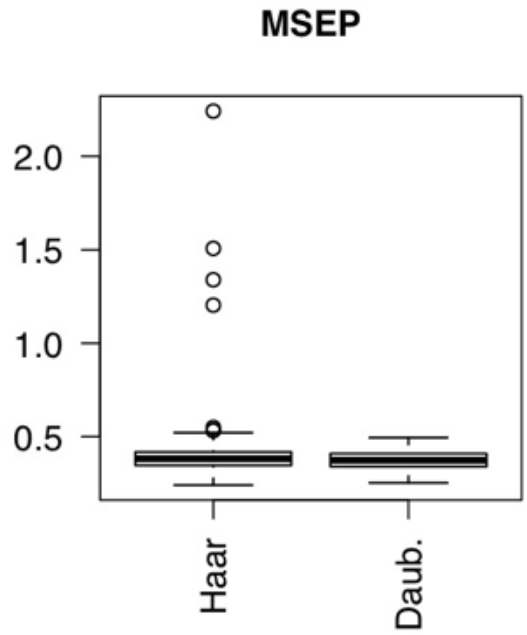


Figure S4. Prediction of Daily Feed Intake. Boxplot of the preprocessing methods considered over 100 resampling replicates, in the model with both metabolomic data and breed information, on preprocessed data with Haar wavelet transformation (Haar) and Daubechies wavelet (Daub.). (A) Mean Square Error of Prediction (MSEP)

Figure S5. Comparison of data pre-processing on prediction errors, for Model 1 (metabolomic data alone as covariates): raw data, Daubechies wavelets, and Haar wavelets.

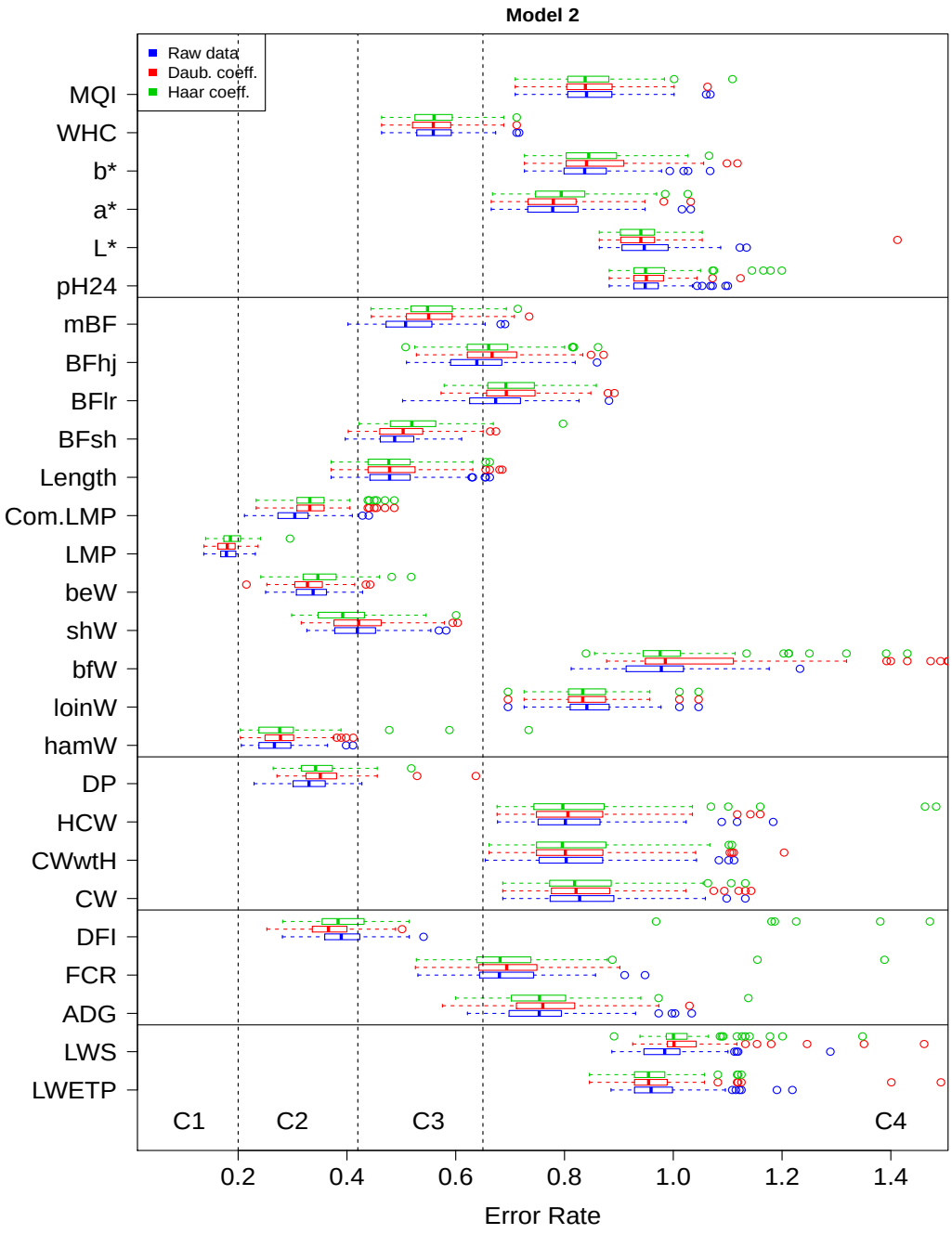


Figure S6. Comparison of data pre-processing on prediction errors, for Model 2 (metabolomic data and breed as covariables): raw data, Daubechies wavelets, and Haar wavelets. Same as Figure 4 in main text.

Figure S7. Comparison of data pre-processing on prediction errors, for Model 3 (metabolomic data, breed and batch as covariables): raw data, Daubechies wavelets, and Haar wavelets.

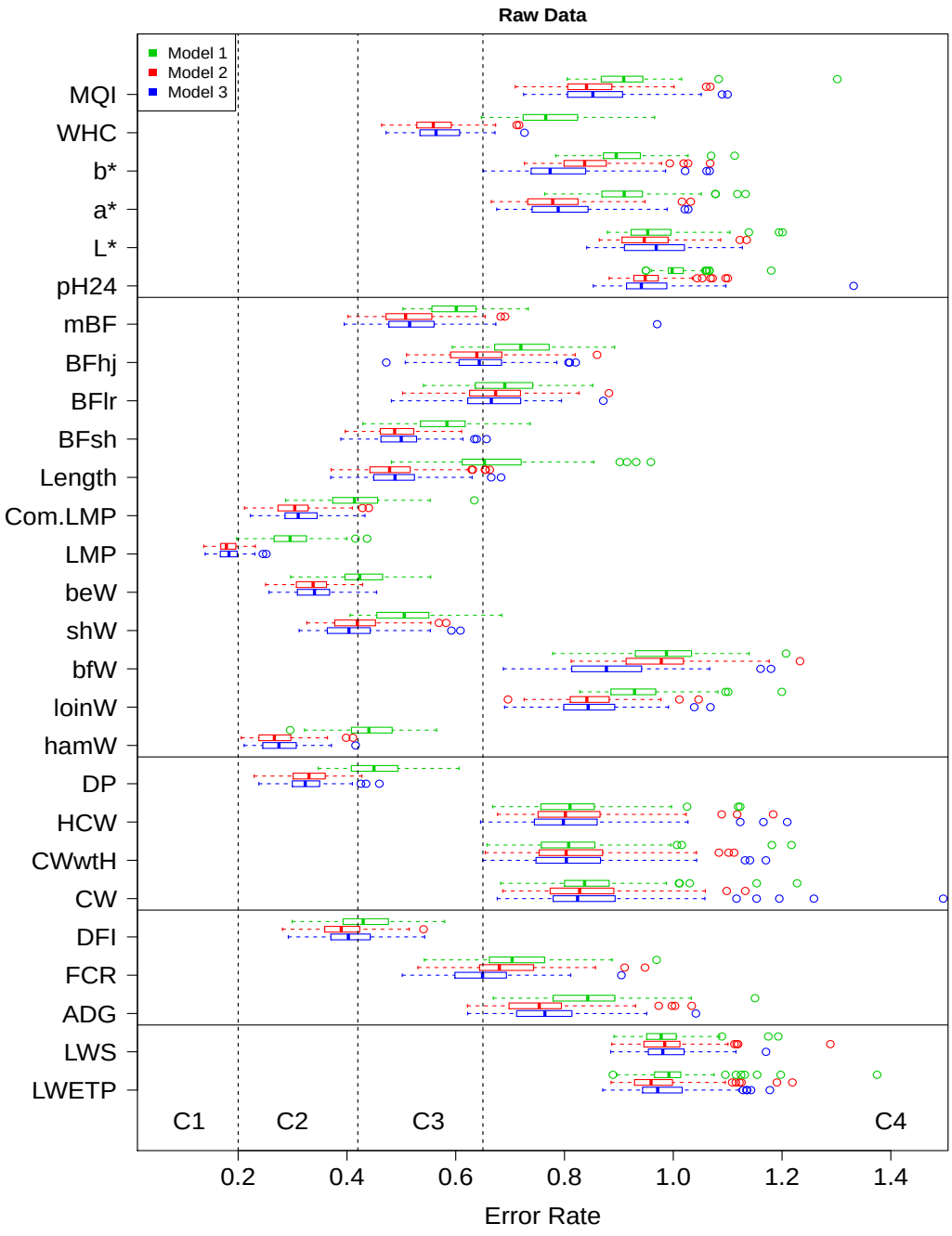


Figure S8. Comparison of model performances on prediction errors, for raw data. Model 1: metabolomic spectra; Model 2: metabolomic spectra and breed; Model 3: metabolomic spectra, breed and batch as covariables.

Figure S9. Comparison of model performances on prediction errors, for pre-processed data (with Haar wavelets). Model1: metabolomic spectra; Model 2: metabolomic spectra and breed; Model 3: metabolomic spectra, breed and batch as covariables.

Figure S10. Comparison of model performances on prediction errors, for pre-processed data (with Daubechies wavelets). Model1: metabolomic spectra; Model 2: metabolomic spectra and breed; Model 3: metabolomic spectra, breed and batch as covariables.

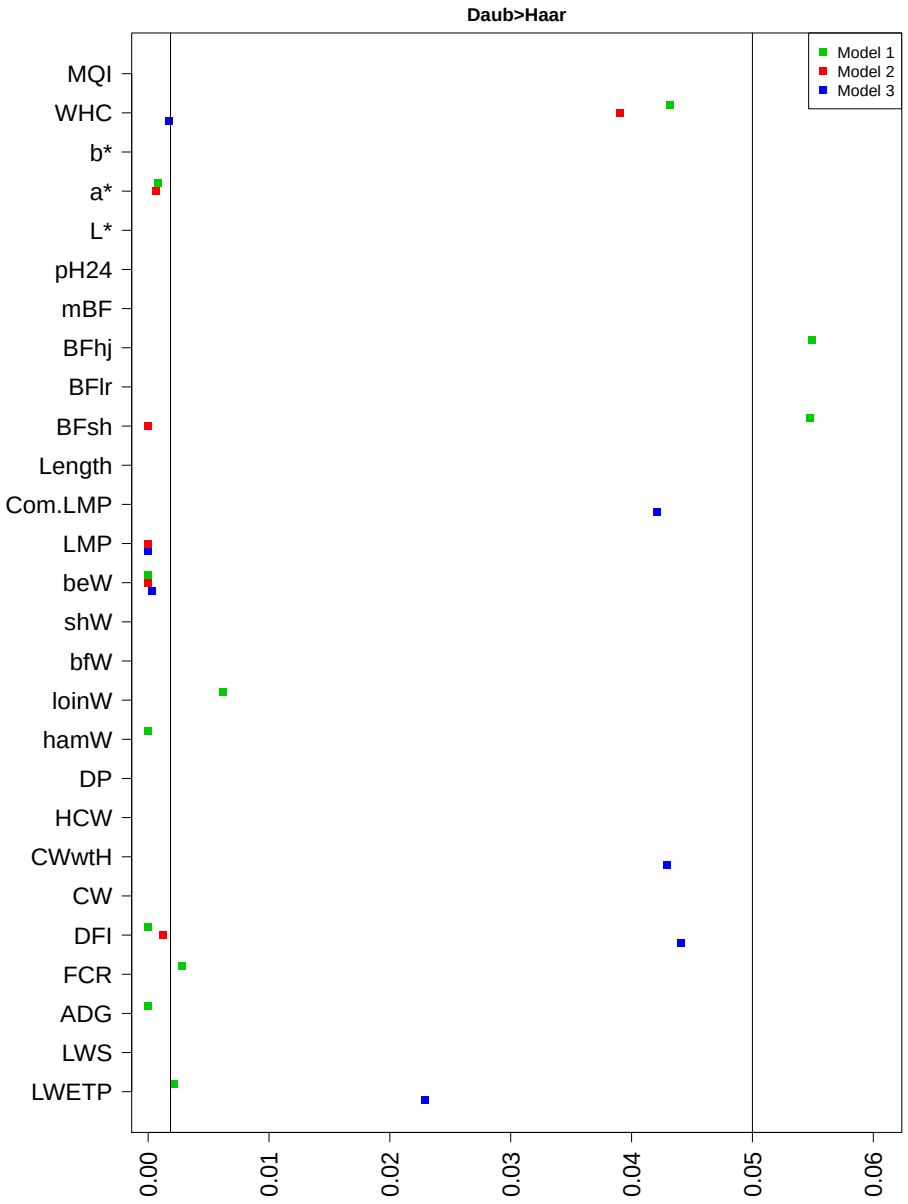


Figure S11. P-values of unilateral paired t-tests for the null hypothesis that the MSE obtained with Daubechies wavelet preprocessing is smaller than the ones with the Haar basis. The Bonferroni correction for a global type I error of 5% is materialized on the graph.

In summary, Daubechies is preferable to Haar in 5 cases for Model 1, 5 cases for Model 2 and 3 cases for Model 3.

Figure S12. P-values of unilateral paired t-tests for the null hypothesis that the MSE obtained with Daubechies wavelet preprocessing is larger than the ones with the Haar basis. The Bonferroni correction for a global type I error of 5% is materialized on the graph.

In summary, Haar is preferable to Daubechies in 5 cases for Model 1, never for Model 2, and in 1 case for Model 3.

	LW	LR	PI
LWETP	111.82±4.17	111.59±3.82	109.3±5.16
LWS	107.56±4.22	107.37±3.74	106.85±5.06
ADG	973.28±89.15	966.69±82.07	872.5±87.24
FCR	2.68±0.2	2.81±0.21	2.46±0.17
DFI	2.61±0.25	2.71±0.22	2.14±0.18
CW	84.23±3.56	83.1±3.23	87.49±4.32
CWwtH	79.35±3.41	78.41±3.19	82.98±4.24
HCW	39.53±1.76	39.1±1.68	41.43±2.16
DP	78.31±1.28	77.39±1.27	81.88±1.27
hamW	9.57±0.51	9.35±0.49	11.48±0.63
loinW	5.1±0.5	5.24±0.45	4.7±0.48
bfW	9.29±0.57	9.11±0.53	9.37±0.62
shW	10.77±0.72	10.25±0.74	12.48±0.89
beW	3.47±0.61	3.85±0.6	2.07±0.4
LMP	55.69±2.69	53.26±2.82	65.38±1.92
Com.LMP	58.12±2.49	55.41±2.87	64.42±1.88
Length	1015.38±26.49	1029.68±27.49	958.35±27.73
BFsh	16.83±3.51	17.31±3.37	9.63±2.92
BFlr	18.32±3.16	19.06±3.27	13.82±2.63
BFhj	35.37±4.22	34.2±3.94	27.89±4.07
mBF	23.51±3.22	23.52±3.05	17.11±2.7
pH24	5.76±0.2	5.7±0.18	5.66±0.15
L*	50.45±3.93	51.1±3.49	53.12±3.86
a*	9.28±1.58	9.4±1.38	11.34±1.93
b*	5.16±1.6	5.41±1.32	6.86±1.69
WHC	14.54±6.24	12.87±6.65	2.12±2.38
MQI	87.46±3.03	86.43±2.77	84.39±2.07

Table S1. Phenotypic data. For each breed, the mean and standard deviation of all phenotypes are given.

		Model 1		Model 2		Model 3	
		Raw Data	Daubechies Coefficients	Raw Data	Daubechies Coefficients	Raw Data	Daubechies Coefficients
Body weights	LWETP	1.00±0.06	1.00±0.06	0.96±0.07	0.95±0.05	0.97±0.06	0.97±0.06
	LWS	0.98±0.05	1.00±0.05	0.98±0.06	1.00±0.16	0.98±0.05	0.98±0.09
Growth traits	ADG	0.84±0.09	0.80±0.09	0.75±0.08	0.76±0.11	0.76±0.08	0.75±0.07
	FCR	0.70±0.08	0.67±0.07	0.68±0.08	0.69±0.07	0.65±0.07	0.66±0.07
	DFI	0.43±0.06	0.40±0.06	0.39±0.05	0.37±0.05	0.40±0.05	0.40±0.07
Carcass weights	CW	0.84±0.09	0.86±0.08	0.83±0.09	0.82±0.09	0.82±0.12	0.82±0.09
	CWwtH	0.81±0.09	0.83±0.08	0.80±0.09	0.80±0.13	0.80±0.10	0.79±0.10
	HCW	0.81±0.09	0.85±0.09	0.80±0.09	0.81±8.63	0.80±0.10	0.79±0.23
	DP	0.45±0.06	0.44±0.09	0.33±0.04	0.35±0.06	0.32±0.04	0.32±0.13
Carcass composition	hamW	0.44±0.06	0.41±0.05	0.27±0.04	0.28±0.04	0.27±0.04	0.28±0.37
	loinW	0.93±0.06	0.90±0.06	0.84±0.06	0.83±0.06	0.84±0.08	0.84±0.07
	bfW	0.98±0.08	0.96±0.07	0.98±0.08	0.98±4.74	0.88±0.08	0.90±0.015
	shW	0.51±0.06	0.51±0.07	0.42±0.05	0.42±2.02	0.40±0.06	0.40±0.06
	beW	0.42±0.05	0.40±0.05	0.34±0.04	0.33±0.04	0.34±0.04	0.36±0.04
	LMP	0.30±0.04	0.29±0.04	0.18±0.02	0.18±0.02	0.18±0.02	0.19±0.03
	Com.LMP	0.41±0.06	0.41±0.06	0.30±0.05	0.33±0.05	0.31±0.04	0.35±0.05
	Length	0.65±0.09	0.65±0.09	0.48±0.06	0.48±0.07	0.49±0.06	0.49±0.07
	BFsh	0.58±0.06	0.56±0.06	0.49±0.05	0.50±0.06	0.50±0.05	0.52±0.06
	BFlr	0.69±0.07	0.70±0.07	0.67±0.07	0.69±0.07	0.67±0.07	0.69±0.07
	BFhj	0.72±0.07	0.68±0.07	0.64±0.07	0.67±0.07	0.64±0.07	0.66±0.07
	mBF	0.60±0.06	0.58±0.06	0.51±0.06	0.55±0.06	0.52±0.07	0.57±0.06
Meat quality	pH24	1.00±0.03	1.00±0.05	0.95±0.04	0.95±0.04	0.94±0.06	0.94±0.09
	L*	0.95±0.06	0.95±0.05	0.95±0.06	0.94±0.06	0.97±0.06	0.96±0.09
	a*	0.91±0.07	0.91±0.07	0.78±0.07	0.78±0.07	0.79±0.08	0.79±0.09
	b*	0.90±0.06	0.89±0.07	0.84±0.06	0.84±0.08	0.77±0.09	0.80±0.09
	WHC	0.77±0.07	0.77±0.06	0.56±0.05	0.56±0.05	0.56±0.05	0.56±0.05
	MQI	0.91±0. 06	0.92±0.05	0.84±0.07	0.84±0.06	0.85±0.07	0.85±0.12

Table 1: Results for Mean Square Error of Prediction (MSEP) on the raw data and the transformed data (with Daubechies) for three models: metabolomic data alone (Model 1), metabolomic + breed (Model 2) and metabolomic + breed + batch (Model 3). The results are presented as median±standard deviation over 100 replicates.

2.3 Pour aller plus loin

L'article présenté en Section 2.2 s'est focalisé sur trois méthodes : le Lasso appliqué sur les données brutes, et le Lasso appliqué sur les coefficients d'ondelettes obtenus par une transformée dans la base de Daubechies ou de Haar. L'estimation du Lasso étant biaisé (Zhang and Hunag, 2008), cette méthode a été utilisée en tant que méthode d'estimation de support et non d'estimation des coefficients. Les coefficients sont estimés à l'aide de l'estimateur des moindres carrés dans le modèle $Y = X_{\hat{S}}\beta_{\hat{S}} + \epsilon$ où $\hat{S}$ est une estimation du support de β obtenue par la méthode Lasso. La validation croisée a donc été réalisée sur la base de cette nouvelle estimation.

La faible différence entre le pouvoir prédictif des données brutes et celui des coefficients d'ondelettes pour le modèle (12b) ou le modèle (12c) pourrait être partiellement expliquée par la Figure 4. Cette figure compare les erreurs de prédictions entre le modèle (12b) et un modèle dans lequel seul l'effet race est pris en compte :

$$\text{phénotype} = \text{intercept} + \text{race} + \text{bruit}. \tag{13}$$

On observe que l'apport des données métabolomiques dans le modèle (12b) en terme d'amélioration de la qualité de prédiction est limité pour la plupart des phénotypes. Ce phénomène signifie que la connaissance du métabolome apporte très peu d'informations supplémentaires par rapport à la connaissance de la race de l'animal pour la prédiction de la plupart des phénotypes. Le phénomène est confirmé par une analyse intra-race présentée en Figure 5 dans laquelle on observe que la prédiction chute considérablement, ici pour la race Large White type femelle. Une autre raison est ici envisageable à celle d'une absence de relation entre le phénome et le métabolome : le pouvoir prédictif diminue puisque le nombre d'observations est plus réduit mais avec autant de variables.
Certaines prédictions sont toutefois améliorées par l'apport des données métabolomiques, comme c'est le cas pour la consommation moyenne journalière DFI et pour deux taux de muscle LMP et Com.LMP, cf. Figure 4. Pour ces phénotypes, l'utilisation conjointe des données métabolomiques et de la race de l'animal permet d'obtenir une qualité de prédiction acceptable. De plus, l'acquisition des données métabolomiques est relativement peu coûteuse et nécessite une simple prise de sang sur l'animal ; la bonne prédiction de phénotypes tel que le taux de muscle présente donc un réel intérêt économique pour la filière porcine.

Les résultats présentés sont basés sur l'utilisation de la méthode Lasso. Toutefois, comme mentionné dans l'article présenté en Section 2.2, d'autres méthodes ont été exploitées comme la méthode PLS, la sPLS ou les forêts aléatoires ; cependant ces méthodes donnant des résultats similaires à la méthode Lasso, elles n'ont pas été étudiées plus avant. On peut observer les différents résultats obtenus en termes d'erreurs de prédictions sur la Figure 3 pour le phénotype DFI et le modèle (12a) appliqué aux données brutes.

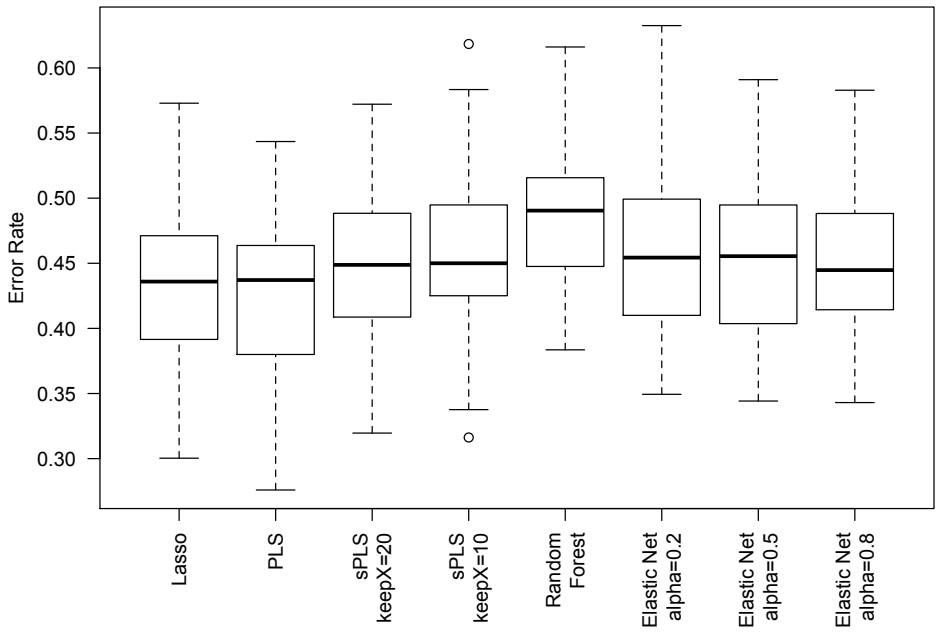


FIGURE 3 – **Erreurs de prédiction concernant le phénotype DFI pour le modèle** (12a) **appliqué au données brutes. Le nombre de directions PLS et sPLS a été choisi par validation croisée sur l'échantillon d'apprentissage. Le coefficients** α **de l'Elastic Net définit la pénalité comme :** $(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1$**.**

2.4 Conclusion

Les résultats présentés dans cette partie proviennent d'une analyse à grande échelle sur des animaux domestiques visant à prédire des phénotypes de productions à partir de profils métabolomiques. Une unique prise de sang a été effectuée sur les animaux à un poids d'environ 60kg ; cependant une certaine variabilité sur le poids ainsi que l'age des animaux au moment de la prise de sang est présente. Cette variabilité combinée au plan d'expérience très déséquilibré nous montre les limitations de cette étude.
Une transformée préalable des profils métabolomiques à l'aide d'outils d'analyse du signal, plus particulièrement les ondelettes, a été proposée. Une amélioration du pouvoir prédictif du métabolome après transformée en ondelettes est visible dans certains cas. L'apport de la transformation en ondelettes est relativement faible voire décevant, contrairement à ce qui était attendu au vu d'études similaires mais non publiées (Martin, Besse, Déjean, Villa-Vialaneix, communications personnelles). L'article présenté ayant un fort intérêt appliqué, seule la méthode de sélection de variables Lasso a été proposée.

L'objectif de cette partie était d'évaluer le pouvoir prédictif de profils métabolomiques de type RMN sur des phénotypes de production utilisés en routine pour l'évaluation des animaux dans la filière porcine. Une étude de cette ampleur est une première. Elle a permis

de conclure que certains phénotypes de production peuvent être prédits à l'aide de profils métabolomiques. Un raffinement du protocole expérimental doit maintenant être réfléchi, notamment sur le moment optimal des prises de sang. Une comparaison plus approfondie de méthodes statistiques de prédiction sera alors nécessaire.

Dans l'objectif d'expliciter la relation entre un phénotype de production et les profils métabolomiques, des méthodes de sélection de variables plus stables que la méthode Lasso sont à envisager. La partie suivante propose de nouvelles méthodes qui ont été développées au cours de cette thèse et qui répondent à ce problème. Ces méthodes seront appliquées sur les données brutes afin de faciliter l'interprétation biologique des variables sélectionnées.

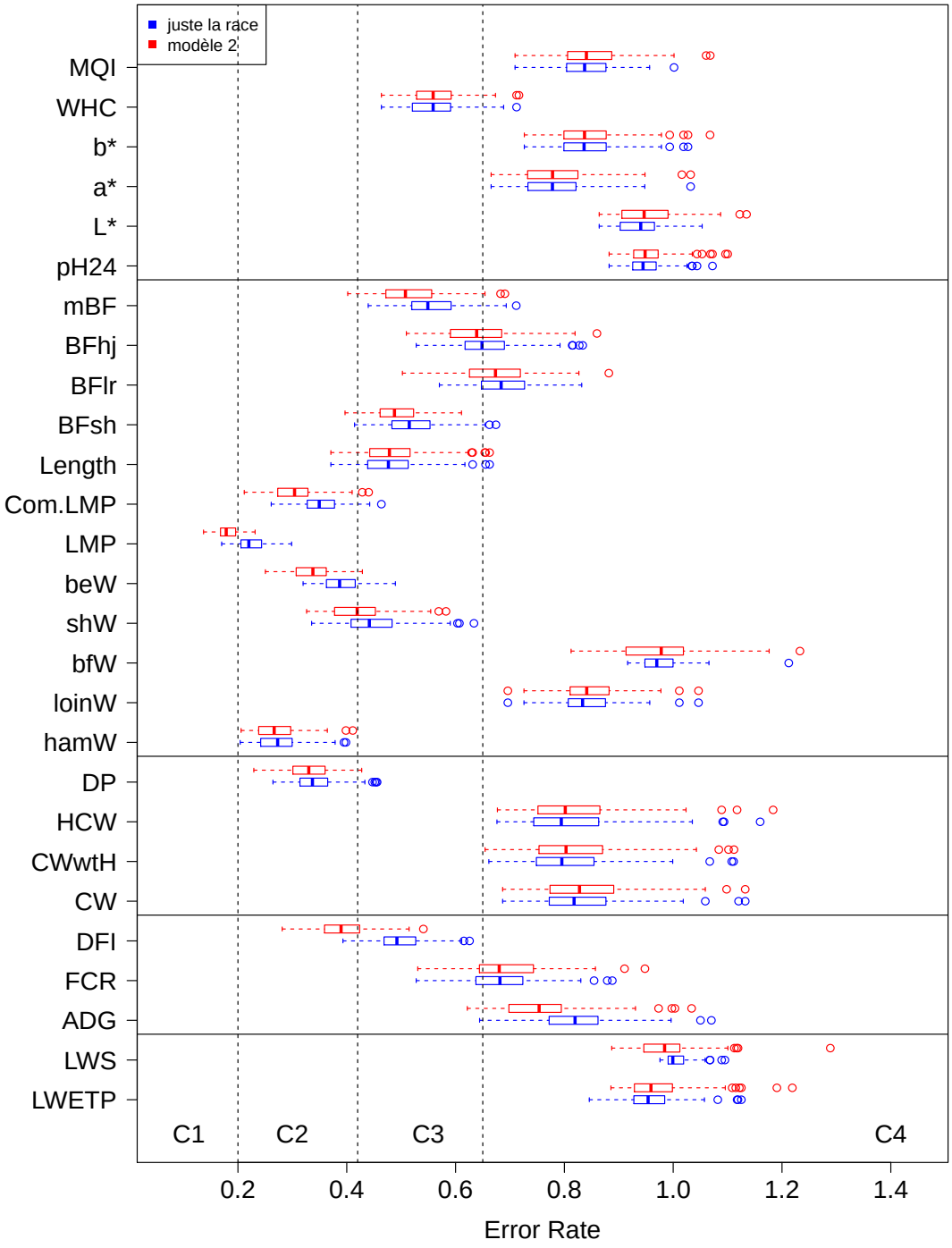


FIGURE 4 – **Prédiction des 27 phénotypes avec la méthode Lasso sur les données brutes, pour le modèle 2 qui considère la race et le métabolome- modèle (12b)- et un modèle ne considérant que la race (13).**

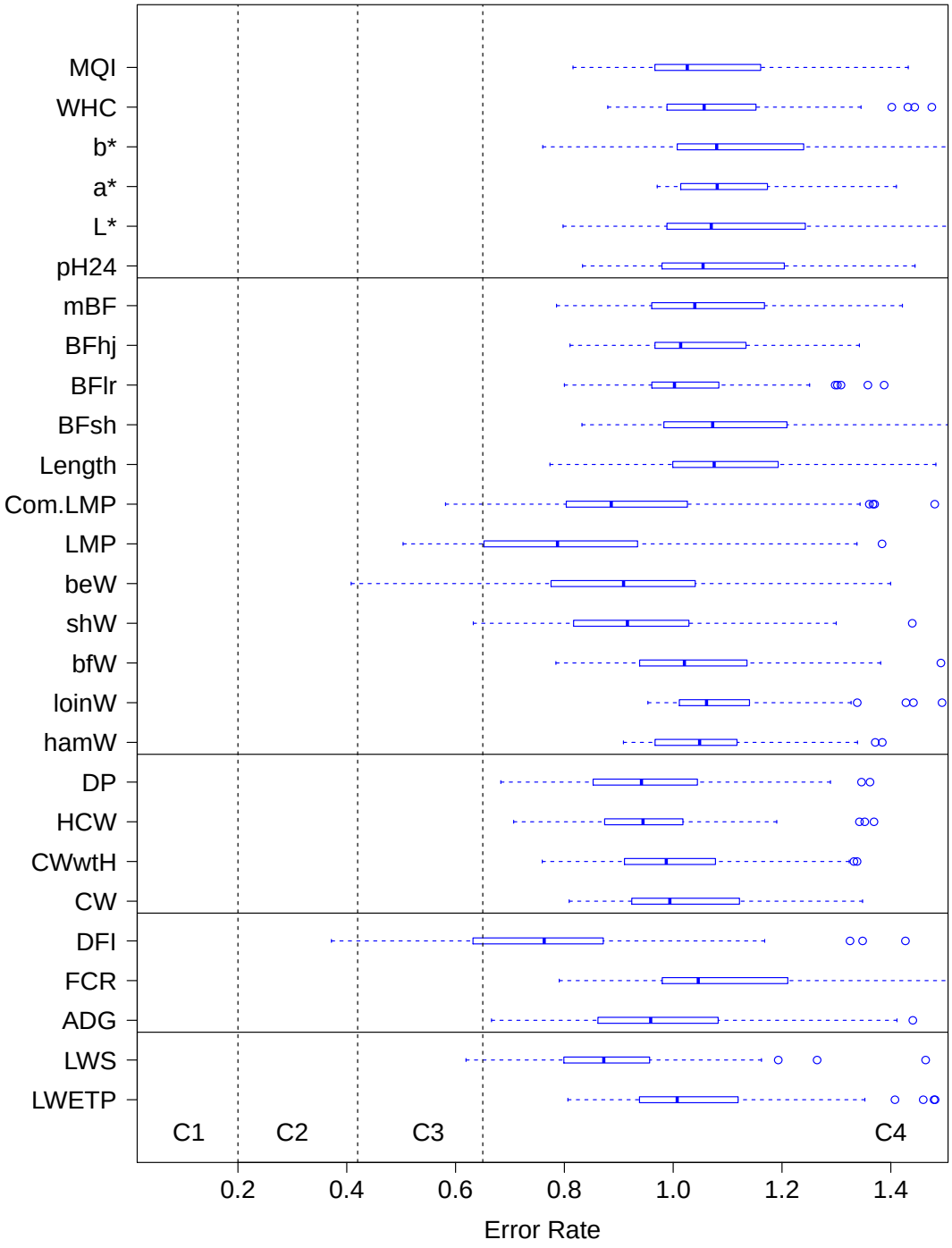


FIGURE 5 – **Prédiction des 27 phénotypes avec la méthode Lasso sur les données brutes, pour le modèle 1 en ne considérant que les individus Large White type femelle.**

3 Sélection de variables dans un modèle linéaire : tests d'hypothèses multiples

3.1 Motivations

Dans un objectif d'explication d'un phénomène biologique, la sélection de variables est la méthode la plus naturelle. Souvent utilisée dans un modèle linéaire, la sélection est appliquée à l'aide de nombreuses méthodes existantes, comme le Lasso, l'adaptive Lasso, le Bolasso ou encore la procédure FDR qui est très prisée des biologistes. La meilleure méthode est celle qui explique entièrement le phénomène sous-jacent (sélectionne toutes les variables pertinentes) sans pour autant insérer trop de faux positifs (variables sélectionnées à tort). Notre objectif est tout d'abord de tester quelques méthodes existantes sur un exemple qui se rapproche de nos données réelles, et de par la suite envisager de nouvelles méthodes de sélection de variables performantes en grande dimension.
Afin de juger du potentiel de ces méthodes en pratique, considérons une simple simulation de grande dimension dans laquelle $n = 100$, $p = 600$ et le nombre de vraies variables est $k_0 = 11$, ce qui en fait un cas à la limite de la très grande dimension : $\frac{k}{n} ln(\frac{p}{k}) = 0.44$ (cf. Section 1.3). Les variables $X_2, \dots, X_p$ sont simulés comme $p-1$ vecteurs gaussien i.i.d. que l'on normalise $\sum_{i=1}^{n} X_{ij}^2 = 1, \forall 2 \leq j \leq p$, et X_1 est une colonne de $1/\sqrt{n}$ -l'intercept-.
La variable Y est simulée comme $Y = \beta_{i_1} X_{i_1} + \dots + \beta_{i_{k_0-1}} X_{i_{k_0-1}} + \epsilon$, où ϵ est un bruit gaussien centré réduit, $J = \{1, i_1, \dots, i_{k_0-1}\} \subset \{1, \dots, p\}$ et $\beta_J = 10$.
La Table 3 montre les résultats des méthodes sus-citées sur 500 simulations de ce simple cadre. La pénalité du Lasso et de ces variantes a été choisie par 10-validation croisée. La procédure FDR est utilisée pour un niveau de contrôle de faux positifs de $q = 0.1$ et $q = 0.05$.

	Idéal	FDR		Lasso	Bolasso	adLasso
		q=0.1	q=0.05			
Egalité	1.00	0.00	0.00	0.00	0.25	0.01
Incl.	11.00	3.33	3.02	17.97	13.24	17.45
C. incl.	11.00	3.33	3.02	10.99	10.99	10.97
MSE	0.00	6.34	6.69	0.31	0.20	0.31

TABLE 3 – **Résultats de** 500 **simulations pour un modèle dans lequel** $n = 100, p = 600, k_0 = 11, \beta_J = 10$**. La première ligne "Egalité" donne le pourcentage de fois où** $\hat{J} = J$**. "Incl." donne la moyenne du nombre de variables sélectionnées et "C. incl." celle du nombre de variables pertinentes sélectionnées. Le MSE est obtenu par moyenne sur toutes les simulations :** $MSE = \sum_{i=1}^{n} (\hat{Y}_i - (X\beta_J)_i)^2/n$**, où** $\hat{Y} = X\hat{\beta}$**, et où** $\hat{\beta}$ **est une estimation de** β **avec des coefficients non nuls seulement sur** $\hat{J}$**.**

Les méthodes testées dans le cadre de cette simulation ne sont pas satisfaisantes en

terme de sélection de variables : le Lasso et ses extensions sélectionnent trop de variables non pertinentes, et la procédure FDR sous estime le nombre de vraies variables sans toutefois prendre en compte de faux positifs.
Partant de ce constat, nous avons développé des méthodes de sélection de variables puissantes basées sur des tests d'hypothèses multiples et donnant de très bon résultats en simulation dans le cas où le nombre de variables explicatives p est plus petit que le nombre d'observations n, mais aussi dans le cas où $p > n$. Ces méthodes sont basées sur une procédure de Baraud et al. (2003) et elles ne contiennent pas de paramètres à optimiser qui influencent fortement les résultats comme c'est le cas pour les méthodes pénalisées (Lasso ou variantes par exemple), le seul paramètre est le niveau du test que l'on fixe au préalable, comme c'est le cas pour la procédure FDR.

3.2 Article - Tests d'hypothèses multiples pour la sélection de variables

Résumé De nombreuses méthodes ont été développées pour estimer l'ensemble des vraies variables d'un modèle linéaire parcimonieux $Y = X\beta + \epsilon$ dans lequel la dimension p de β peut être beaucoup plus grande que la longueur n du vecteur d'observations. Nous proposons deux nouvelles méthodes de sélection de variables basées sur des tests d'hypothèses multiples, une méthode concerne la sélection ordonnée et une autre la sélection non ordonnée. Nos procédures sont inspirées de la procédure de tests multiples introduit par Baraud et al. (2003). Les nouvelles procédures sont puissantes sous certaines conditions sur le signal $X\beta$ et leurs propriétés sont non asymptotiques. Ces procédures donnent de meilleurs résultats que la procédure FDR et le Lasso, en petite dimension ($p < n$) mais aussi en grande dimension ($p \geq n$).

Article soumis

Multiple Hypotheses Testing For Variable Selection

Florian Rohart

April 2012

Abstract

Many methods have been developed to estimate the set of relevant variables in a sparse linear model $Y = X\beta + \epsilon$ where the dimension p of β can be much higher than the length n of Y. Here we propose two new methods based on multiple hypotheses testing, either for ordered or non-ordered variable selection. Our procedures are inspired by the testing procedure proposed by Baraud et al. (2003). The new procedures are proved to be powerful under some conditions on the signal and their properties are non asymptotic. They gave better results in estimating the set of relevant variables than both the False Discovery Rate (FDR) and the Lasso, both in the common case ($p < n$) and in the high-dimensional case ($p \geq n$).

1 Introduction

Recent technologies have provided scientists with very high-dimensional data. This is especially the case in biology with high-throughput DNA/RNA chips. Unravelling the relevant variables -genes for example- underlying an observation is a well known problem in statistics and is still one of the current major challenges. Indeed, with a large number of variables there is often a desire to select a smaller subset that not only fits almost as well as the full set of variables, but also contains the most important ones for a prediction purpose. Discovering the relevant variables leads to higher prediction accuracy, an important criterion in variable selection.

Many methods have been developed to estimate the set of relevant variables in the linear model $Y = X\beta + \epsilon$ where the dimension p of β can be much higher than the length n of Y. Most of these methods are based on a penalized criterion. The mostly known is probably the Lasso that has been presented by Tibshirani (1996); l^1 penalization of the least squares estimate which shrinks to zero some irrelevant coefficients, hence an estimation of the set of relevant variables. A lot of studies have been conducted on the Lasso and many results are available (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006; Bunea et al., 2007; Wainwright, 2009). The Lasso has several variants such as an adaptative Lasso (Huang et al., 2008), a bootstrap Lasso (Bach, 2009) or a Group Lasso (Chesneau and Hebiri, 2008). A l^1 penalization has also been used in the Sparse-PLS, which induces a limited number of variables in each PLS direction; see Tenenhaus (1998) for an introduction on PLS, and Lê Cao et al. (2008) for further details on Sparse-PLS. Other kinds of penalization have also been used, such as the Akaike Information Criterion

(AIC) or the Bayesian Information Criterion (BIC), two methods based on the logarithm of the likelihood penalized by the number of variables included in the model. Despite that the major portion of model selection methods was developed to perform in low dimension, some of them apply in the high-dimensional case. There is still some others that were actually developed to be powerful when p is higher than n, such as the Dantzig selector (Candes and Tao, 2007). Yet, a recent paper shows that under a sparsity condition on the linear model, the Dantzig selector and the Lasso exhibit similar behavior (Bickel et al., 2009). Nevertheless, penalization criterion is not the only way to perform model selection. For instance, the False Discovery Rate (FDR) procedure, developed in the context of multiple hypotheses testing by Benjamini and Hochberg (1995), was used in variable selection by Bunea et al. (2006). This procedure has been extended to high-dimensional analysis and is presently used in biology for QTL research and transcriptome analysis; a p-value is calculated for each variable X_i from the regression of Y onto that variable and selection is performed through an adjusted threshold. Wasserman and Roeder (2009) proposed a three stages procedure that also uses hypotheses testing, they called it 'the screen-and-clean procedure'. The first stage fits a collection of models through a chosen method -they proposed the Lasso, the marginal regression and the forward stepwise regression-, the second stage selects a model among that collection thanks to cross validation, finally the last step uses hypothesis testing to perform variable selection. The screen-and-clean procedure is consistent under certain conditions.

Most of the selection methods cited above give quite good results when p is lower than n. However, they all have drawbacks that especially appear in a high-dimensional context. For instance, Lasso lacks stability: due to increasing collinearity when $p > n$, only small changes in the data set leads to different sets of selected variables. Moreover, the results of the Lasso, as well as its extensions, depend on a penalty parameter that has to be tuned, which is surely the major drawback. For the screen-and-clean procedure to be efficient, the dataset has to be divided in three, which is not always conceivable when the number of observations is small. As the authors pointed out in their simulation, they obtained better results when the first two steps were both conducted on the same split of the data, leading to the question of usefulness of the data split in practice. Moreover, the variance is assumed to be known in their theoretical results.
This paper deals with the problem of selecting the set of indices of the relevant variables in a sparse linear model when p can be lower or higher than n without data splitting and with unknown variance. We present a new method of variable selection based on multiple hypotheses testing which is stable and free of tuning parameters, except the type I error of the tests which has to be chosen (as for the FDR procedure or any statistical testing).

We consider the regression model:

$$Y = X\beta + \epsilon, \tag{1}$$

where Y is the observation of length n, $X = (X_1, \dots, X_p)$ is the $n \times p$ matrix of p variables, β is an unknown vector of $\mathbb{R}^p$, ϵ a Gaussian vector with i.i.d. components, $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ where I_n is the identity matrix of $\mathbb{R}^n$, and σ some unknown positive quantity. For the

convenience of the reader, one could recall that X_1 is the intercept. We define the support of β, $J = \{j, \beta_j \neq 0\}$ and $|J| = k_0$. A variable X_j is said to be relevant when $\beta_j \neq 0$. We denote $\beta_J = (\beta_j)_{j \in J}$. Let $\mu = E(Y) = X\beta$ and $\mathbb{P}_\mu$ the distribution of Y obeying to model (1).

The aim of this paper is to estimate J, the set of indices of the relevant variables in (1). We distinguish two frameworks. In a first step, we only consider ordered variable selection. We define a powerful procedure for estimating J under some conditions on the signal, either when $p \leq n$ or when $p > n$. These properties are non asymptotic. The procedure is a multiple hypotheses testing method based on the testing procedure developed by Baraud et al. (2003). In a second step, the variables are not assumed to be ordered. We provide a procedure to estimate J when σ is known and another procedure when σ is unknown. The two procedures are proved to be powerful under some conditions on the signal. The properties of the procedures are also non asymptotic.

Let us introduce some notations that will be used throughout this paper. Note $||s||_n^2 = \sum_{i=1}^n s_i^2/n$. Set Π_V the orthogonal projector onto V for all subspace V. $\bar{F}_{D,N}(u)$ denotes the probability for a Fisher variable with D and N degrees of freedom to be larger than u. We denote $\forall\ (x,y) \in \mathbb{R}^{n^2} < x, y >_n = \sum_{i=1}^n x_i y_i/n$, $< x, y >= n < x, y >_n$ and $\forall a \in \mathbb{R}$, $\lfloor a \rfloor$ the integer part of a.

This paper is organized as follow, in Section 2 we present the first procedure to estimate J in the context of ordered variable selection; the non-ordered variable selection is considered in Section 3. A simulation study is provided in Section 4 to compare several variable selection methods. The proofs are given in B.

2 Ordered variable selection

The procedure that will be described in this section is applicable either when $p < n$ or when $p \geq n$. We make the following assumptions :

- A_1: each family $\{X_I, I \subset \{1, \ldots, p\}, |I| = \min(p, n)\}$ is linearly independent,
- A_2: the number of relevant variables verifies $k_0 \leq \min(n-1, p)$.

For all $i \in \{1, \ldots, p\}$, X_i is supposed to have unit variance: $\forall\ i, \frac{1}{n}\sum_{j=1}^n X_{ij}^2 = 1$.

In this section we focus on ordered variables selection, which means that the set of indices of the relevant variables is supposed to be $J = \{1, \ldots, k_0\}$, for some $k_0 \leq \min(n-1, p)$. Hence an estimation of k_0 gives us an estimation of J. This section focuses on the estimation of k_0.

Our procedure is a multiple hypotheses testing method based on the testing procedure developed by Baraud et al. (2003) in the context of linear regression of $Y = f + \epsilon$ where f is an unknown vector of $\mathbb{R}^n$. Let V be a subspace of $\mathbb{R}^n$. They constructed a testing procedure of the null hypothesis "f belongs to V" against the alternative that it does not under no prior assumption on f. Their testing procedure is based on the choice

of a collection $\{S_m, m \in \mathcal{M}\}$ of subspaces of $V^\perp$ and the choice of a collection of levels $\{\alpha_m, m \in \mathcal{M}\}$. They considered for each $m \in \mathcal{M}$ a Fisher test of level α_m to test

$$H_0 : \{f \in V\} \quad \text{against the alternative} \quad H_{1,m} : \{f \in (V + S_m)\backslash V\},$$

and the null hypothesis H_0 is rejected if one of the Fisher test does.

Our procedure consists in applying the procedure proposed by Baraud et al. (2003) on a collection of subspaces $(V_k)_{1 \leq k < \min(n-1,p)}$ to test successively the null hypotheses H_k : $\{\mu \in V_k\}$, for $1 \leq k < \min(n-1, p)$. We stop our procedure as soon as a null hypothesis is accepted.
Set $\forall 1 \leq k < \min(n-1, p), V_k = \text{span}(X_1, \ldots, X_k)$. With this choice of V_k and assumption A_1, we have $\dim(V_k) = k, \forall 1 \leq k < \min(n-1, p)$.
Let k be fixed in $\{1, \ldots, \min(n-1, p) - 1\}$, we define $t^k_{max} = \lfloor \log_2(\min(n-1, p) - k) \rfloor$ and $\mathcal{T}_k = \left\{0, \ldots, t^k_{max}\right\}$.
As done in Baraud et al. (2003), given a collection of levels $\{\alpha_{k,t},\ t \in \mathcal{T}_k\}$ and a collection of linear spaces $\{S_{k,t},\ t \in \mathcal{T}_k\}$ we consider for each $t \in \mathcal{T}_k$ a Fisher test of level $\alpha_{k,t}$ to test the null hypothesis

$$H_k : \{\mu \in V_k\} \qquad \text{against the alternative} \qquad \{\mu \in (V_k + S_{k,t})\backslash V_k\}.$$

The null hypothesis H_k is rejected if at least one of the Fisher tests does. The collection of levels $\{\alpha_{k,t},\ t \in \mathcal{T}_k\}$ is calibrated in order to ensure that the final test H_k is of level α -fixed in $]0, 1[$-, and the collection $\{S_{k,t},\ t \in \mathcal{T}_k\}$ of linear subspaces of $V^\perp$ is defined as follows:
$\forall\ t \in \mathcal{T}_k$,

$$S_{k,t} = \text{span}\left(\Pi_{V_k^\perp}(X_{k+1}), \ldots, \Pi_{V_k^\perp}(X_{k+2^t})\right). \tag{2}$$

Let us introduce some notations that will be used throughout this section. For each $k \in \{1, \ldots, \min(n-1, p) - 1\}$, $t \in \mathcal{T}_k$, we set $V_{k,t} = V_k \oplus S_{k,t}$, and denote $D_{k,t} = 2^t$ and $N_{k,t} = n - (k + 2^t)$ the dimension of $S_{k,t}$ and $V_{k,t}^\perp$ respectively.

As our procedure consists in successively testing the null hypotheses $(H_k)_{1 \leq k < \min(n-1,p)}$ at level α until a null hypothesis is accepted, an estimation of k_0 with our procedure is

$$\hat{k} = \inf\{k \geq 1, H_k \text{ is accepted}\}.$$

The estimated set of indices of the relevant variables is then $\hat{J} = \left\{1, \ldots, \hat{k}\right\}$. Note that if all the null hypotheses $(H_k)_{1 \leq k < \min(n-1,p)}$ are rejected, $\hat{J} = \{1, \ldots, \min(n-1, p)\}$.
Let us recall the definition of the procedure proposed by Baraud et al. (2003) to test the null hypothesis H_k. Set: $\forall \alpha \in]0, 1[$, $\forall 1 \leq k < \min(n-1, p)$,

$$T_{k,\alpha} = \sup_{t \in \mathcal{T}_k} \left\{ \frac{N_{k,t} ||\Pi_{S_{k,t}} Y||_n^2}{D_{k,t} ||Y - \Pi_{V_{k,t}} Y||_n^2} - \bar{F}^{-1}_{D_{k,t}, N_{k,t}}(\alpha_{k,t}) \right\}, \tag{3}$$

where $\{\alpha_{k,t},\ t \in \mathcal{T}_k\}$ is a collection of number in]0,1[such that:

$$\forall \mu \in V_k, \qquad \mathbb{P}_\mu(T_{k,\alpha} > 0) \leq \alpha. \tag{4}$$

The null hypothesis H_k is rejected when $T_{k,\alpha}$ is positive.
They chose the collection $\{\alpha_{k,t}, t \in \mathcal{T}_k\}$ in accordance with one of the two following procedures:

P1. For all $t \in \mathcal{T}_k, \alpha_{k,t} = \alpha_{k,n}$ where $\alpha_{k,n}$ is the α-quantile of the random variable

$$\inf_{t \in \mathcal{T}_k} \bar{F}_{D_{k,t},N_{k,t}} \left\{ \frac{N_{k,t}||\Pi_{S_{k,t}}\epsilon||_n^2}{D_{k,t}||\epsilon - \Pi_{V_{k,t}}\epsilon||_n^2} \right\},$$

P2. The collection $\{\alpha_{k,t}, t \in \mathcal{T}_k\}$ satisfies the inequality

$$\sum_{t \in \mathcal{T}_k} \alpha_{k,t} \leq \alpha.$$

Procedure P1 gives a test H_k of size α whereas procedure P2, which relies on Bonferonni's inequality, only gives a test H_k of level α. Our final multiple testing procedure, which consists in calculating successively $T_{k,\alpha}$ from $k = 1$ until $T_{k,\alpha}$ is negative, is proved to be powerful. An upper bound of the probability to wrongly estimate k_0 is given in the following theorem. Let us first introduce some notations. For $k = 1, \ldots, \min(n-1, p) - 1$, for $\gamma \in]0,1[$ and for all $t \in \mathcal{T}_k$, let $L_t = \log(1/\alpha_{k,t}), L = \log(2/\gamma), m_t = 2\exp(4L_t/N_{k,t})$, and for $u > 0$ let

$$K_t(u) = 1 + 2\sqrt{\frac{u}{N_{k,t}}} + 2m_t\frac{u}{N_{k,t}},$$

$$C_1(k,t) = 2.5(1 + K_t(L_t) \vee m_t)\frac{D_{k,t} + L_t}{N_{k,t}},$$

$$C_2(k,t) = 2.5\sqrt{1 + K_t^2(L)}\left(1 + \sqrt{\frac{D_{k,t}}{N_{k,t}}}\right),$$

$$C_3(k,t) = 2.5\left[\left(\frac{m_t K_t(L)}{2}\right) \vee 5\right]\left(1 + 2\frac{D_{k,t}}{N_{k,t}}\right),$$

Theorem 2.1. *Let Y obey to Model (1). Assume that conditions A_1 and A_2 are verified. We denote by J the set $\{j, \beta_j \neq 0\} = \{1, \ldots, k_0\}$. Let γ and α be fixed in $]0,1[$.*
The testing procedure estimates k_0 by $\hat{k} = inf\{k \geq 1, T_{k,\alpha} \leq 0\}$, where $T_{k,\alpha}$ is defined by (3). Let $\{\alpha_{k,t}, t \in \mathcal{T}_k\}$ be defined according to the procedure P1 or P2.
The following inequality holds for all $\mu \in \mathbb{R}^n$ and for all $k_0 \leq min(n-1, p)$:

$$\mathbb{P}_\mu(\hat{k} > k_0) \leq \alpha. \tag{5}$$

If $\forall k \leq k_0 - 1$ the condition (R_k) holds

$(R_k) : \exists\ t \in \mathcal{T}_k /$

$$\begin{aligned} \left|\left|\Pi_{S_{k,t}}(\mu)\right|\right|_n^2 &\geq C_1(k,t)\left|\left|\Pi_{V_{k,t}^{\perp}}(\mu)\right|\right|_n^2 \\ &+\frac{\sigma^2}{n}\left[C_2(k,t)\sqrt{2^t \log\left(\frac{2k_0}{\alpha_{k,t}\gamma}\right)} + C_3(k,t)\log\left(\frac{2k_0}{\alpha_{k,t}\gamma}\right)\right] \end{aligned}$$

then

$$\mathbb{P}_\mu(\hat{k} < k_0) \leq \gamma, \tag{6}$$

which implies that

$$\mathbb{P}_\mu(\hat{k} \neq k_0) \leq \gamma + \alpha. \tag{7}$$

This result is derived from the result on the power of the multiple testing procedure proposed by Baraud et al. (2003). It is important to note that Theorem 2.1 is non asymptotic.

Comments

1. As mentioned in Baraud et al. (2003), for k fixed, $C_1(k,t)$, $C_2(k,t)$ and $C_3(k,t)$ behave like constants if the following conditions are verified:

   For all $t \in \mathcal{T}_k, \alpha_{k,t} \geq \exp(-N_{k,t}/10),\ \gamma \geq 2\exp(-N_{k,t}/21)$ and the ratio $\dfrac{D_{k,t}+L_{k,t}}{N_{k,t}}$ remains bounded.

   Under these conditions, the following inequalities hold:

$$C_1(k,t) \leq 10\frac{D_{k,t}+\log(1/\alpha_{k,t})}{N_{k,t}},$$

$$C_2(k,t) \leq 5\left(1+\sqrt{\frac{D_{k,t}}{N_{k,t}}}\right),$$

$$C_3(k,t) \leq 12.5\left(1+2\frac{D_{k,t}}{N_{k,t}}\right).$$

2. We say that μ satisfies condition (R) if $\forall k \leq k_0 - 1$, (R_k) holds. According to Theorem 2.1, our procedure is powerful under the condition (R). Assume that $p < n$. A condition on the coefficients β_J underlies in (R) since the projection of Y onto a space spanned by a subset of the family $(X_i)_{1\leq i\leq p}$ depends both on β and on the matrix X. These conditions on β_J explicitly appear when $(X_i)_{1\leq i\leq p}$ is an orthonormal family. Assuming that $(X_i)_{1\leq i\leq p}$ is an orthonormal family, (1) becomes:

$$Y = \underbrace{X_1\beta_1 + .. + X_k\beta_k}_{\in V_k} + \underbrace{X_{k+1}\beta_{k+1} + ... + X_p\beta_p}_{\in V_k^{\perp}} + \epsilon. \tag{8}$$

   With the new decomposition (8), the projection of Y on any subspace $S_{k,t}$ only depends on the coefficients $(\beta_j)_{j\geq k+1}$. Thus the condition (R_k) can be written in a

more explicit form, involving the coefficients $(\beta_j)_{1\leq j\leq p}$. Namely, (R_k) is equivalent to: $\exists\, t \in \mathcal{T}_k /$

$$\beta_{k+1}^2 + .. + \beta_{k+2^t}^2 \quad \geq \quad C_1(k,t) \sum_{j=k+2^t+1}^{p} \beta_j^2 + \frac{\sigma^2}{n}\left[C_2(k,t)\sqrt{2^t \log\left(\frac{2k_0}{\alpha_{k,t}\gamma}\right)} + C_3(k,t)\log\left(\frac{2k_0}{\alpha_{k,t}\gamma}\right)\right].$$

When $k < k_0$, the coefficients $\beta_{k+1}, \ldots, \beta_{k_0}$ are not equal to 0. If for some $t \in \mathcal{T}_k$, the sum $\beta_{k+1}^2 + \cdots + \beta_{k+2^t}^2$ is large enough (namely larger than the right hand of the above equation), then the test will be powerful and the hypotheses H_k will be rejected with high probability.

Results from a simulation study in Section 4 will show the power of our procedure; either when $p < n$ or when $p \geq n$.

3 Non-ordered variable selection

In Section 2 we defined a procedure based on multiple hypotheses testing in order to estimate J, the set of indices of the relevant variables of a sparse linear model (1). As we considered ordered variable selection, the estimation of $J = \{1, \ldots, k_0\}$ was reduced to the estimation of k_0. The present section is dedicated to non-ordered variable selection, so J is not necessarily equal to $\{1, \ldots, k_0\}$. We define here a general two-step procedure to estimate J; the first step orders the variables and the second performs multiple testing. After the first step of the general procedure, the ordered variables will be denoted as $X_{(1)}, \ldots, X_{(p)}$, where $X_{(1)} = X_1$.

The first step of our procedure consists in ordering the variables. It is important to note that the procedure that will be described in this section applies for any possible way to order the variables. However, the order has a strong influence on the final results of our procedure, thus it has to be carefully chosen. Indeed, as we will see throughout this section, the first step is crucial; the ability to estimate J with our procedure depends on the ability to get the relevant variables in the first places, hence on the way to order the variables. In this paper, we considered two ways to order $(X_i)_{2\leq i\leq p}$ taking into account the observation Y.

1. Variables ordered by increasing p-values: when $p < n$, a p-value is calculated for each variable from the test of nullity of the coefficient associated to this variable and the variables are sorted by increasing p-values. When $p \geq n$, a p-value is calculated for each variable using the decomposition of Y onto that variable.

2. The second method that we propose orders the variables with the Bolasso technique, introduced by Bach (2009). It is a bootstrapped version of the Lasso which improves its stability: several independent bootstrap samples are generated and the Lasso is performed on each of them. This approach is proved to make the irrelevant variables asymptotically disappear. A variable X_i is selected by the Bolasso technique at a given penalty if X_i is selected in each bootstrap sample at the same penalty. To avoid the use of a penalty, we set the first ordered variable of the family $(X_i)_{2\leq i\leq p}$ to be the first one to be selected by the Bolasso technique from a decreasing penalty; and so on for the other variables. We proceed by dichotomy to order the variables.

The first method has been considered since it is often used in practice, in particular in the False Discovery Rate procedure (Benjamini and Hochberg, 1995) and the Marginal Regression (Wasserman and Roeder, 2009). It is the one requiring less computational time, but as shown in Section 4, the Bolasso technique gives better results and since the results strongly depends on the ordering on the variables, the Bolasso technique should be preferred.

From now on, we assume that we could be in a high dimensional case and that both assumptions A1-A2 of Section 2 are verified. We introduce here an event that will be useful in the following of this section:

$$A_k = \{\{(1), \ldots, (k)\} = J\}\,. \tag{9}$$

On the event A_k, the set of the k first ordered variables corresponds to the set J of the relevant variables. The second step of the general procedure consists in testing successively the null hypothesis:

$$\hat{H}_k : \ \left\{\mu \in \text{span}(X_{(1)}, \ldots, X_{(k)})\right\} \text{ against the alternative that it does not.} \tag{10}$$

The procedure stops when the null hypothesis is accepted:

$$\mathring{k} = \inf\left\{k \geq 1, \hat{H}_k \text{ is accepted}\right\}.$$

We estimate the set J of relevant variables by

$$\hat{J} = \left\{(1), \ldots, (\mathring{k})\right\}.$$

Note that this is not a simple generalization of the procedure proposed in Section 2 since $\text{span}(X_{(1)}, \ldots, X_{(k)})$ are random spaces depending on the observation Y which have been used in the first step to order the family $(X_i)_{2\leq i\leq p}$. The same observation Y will be used in the second step to perform the multiple testing procedure. A simple generalization of Section 2 could have been constructed from splitting the data in two sets: the first set being used to order the variables, the multiple testing procedure being performed on the second set. Remark that such splitting is the essence of the 'clean-and-screan' procedure of Wasserman and Roeder (2009).

For the sake of understanding, we first deal with the case where σ is known in order to propose a multiple testing procedure.

3.1 Non-ordered variable selection with known variance

In this section, we define a procedure called Procedure 'A' under the assumption that the variance σ^2 is known. Assume that the first step of Procedure 'A' has already been done: variables have been ordered. The second step is a testing procedure that will be described in the following. As in the previous section, we test successively the null hypotheses $\hat{H}_k$ for $1 \leq k < \min(n-1, p)$ until a null hypothesis is accepted.

Let us adapt the notation of Section 2 to this section: we first recall that $\forall\ 1 \leq k < \min(n-1,p)$, $t^k_{max} = \lfloor \log_2(\min(n-1,p) - k) \rfloor$, $\mathcal{T}_k = \left\{0, \dots, t^k_{max}\right\}$. We define $V_{(k)} = \text{span}(X_{(1)}, \dots, X_{(k)})$ and $\forall\ t \in \mathcal{T}_k$, $S_{(k),(t)} = \text{span}\left(\Pi_{V^\perp_{(k)}}(X_{(k+1)}), \dots, \Pi_{V^\perp_{(k)}}(X_{(k+2^t)})\right)$. With the definition of $S_{(k),(t)}$, we have $\dim(S_{(k),(t)}) = D_{k,t} = 2^t$. Let us denote $V_{(k),(t)} = V_{(k)} \oplus S_{(k),(t)}$.
For all $t \in \mathcal{T}_k$, our aim is to test

$$\hat{H}_k : \left\{\mu \in V_{(k)}\right\} \qquad \text{against the alternative} \qquad \left\{\mu \in (V_{(k)} + S_{(k),(t)}) \backslash V_{(k)}\right\}. \tag{11}$$

Since the variance is assumed to be known, we introduce for all $1 \leq k < \min(n-1,p)$ and for all $t \in \mathcal{T}_k$,

$$U_{k,t} = \frac{||\Pi_{S_{(k),(t)}} Y||^2_n}{\sigma^2}.$$

We introduce a multiple testing procedure that relies on the statistics $\{U_{k,t}, t \in \mathcal{T}_k\}$.

Since the spaces $\left\{S_{(k),(t)}, t \in \mathcal{T}_k\right\}$ are random and depend on Y as mentioned before, we first provide a stochastic upper bound for the statistics $U_{k,t}$ in order to define the multiple testing procedure.

Let $\epsilon' \sim \mathcal{N}_n(0, \sigma^2 I_n)$. For all $1 \leq k < \min(n-1,p)$, we define a permutation σ^k_1 of $\{1, \dots, p\}$:

$\sigma^k_1(j) = (j)$ for all $j \in \{1, \dots, k\}$.
For $j \in \{k+1, \dots, p\}$, set $X^{(j)}_i = \Pi_{\text{span}(X_{\sigma^k_1(1)}, \dots, X_{\sigma^k_1(j-1)})^\perp}(X_i)$ for all $1 \leq i \leq p$ and define $\sigma^k_1(j) = \underset{i \in \{1,\dots,p\}}{\text{argmax}} \left|\left|\Pi_{\text{span}(X^{(j)}_i)}(\epsilon')\right|\right|^2_n$.

Set for all $1 \leq k < \min(n-1,p)$ and for all $t \in \mathcal{T}_k$,

$$U^1_{k,t} = \frac{||\Pi_{S_{(k),\sigma^k_1(t)}} \epsilon'||^2_n}{\sigma^2},$$

where $S_{(k),\sigma^k_1(t)} = \text{span}\left(\Pi_{V^\perp_{(k)}}(X_{\sigma^k_1(k+1)}), \dots, \Pi_{V^\perp_{(k)}}(X_{\sigma^k_1(k+2^t)})\right)$.

Note that the distribution of $U^1_{k,t}$ only depends on the design matrix X, and can therefore be simulated for a given matrix X.

Lemma 3.1. *Let $1 \leq k < min(n-1, p)$ and $t \in \mathcal{T}_k$.*
We define $A_k = \left\{\left\{X_{(1)}, \ldots, X_{(k)}\right\} = \{X_j, j \in J\}\right\}$. For all $x > 0$ we have

$$\mathbb{P}\left((U_{k,t} > x) \cap A_k\right) \leq \mathbb{P}\left(U^1_{k,t} > x\right).$$

Let $\overline{U^1_{k,t}}(u)$ denote the probability for the statistic $U^1_{k,t}$ to be larger than u.
Set $\forall \alpha \in]0,1[$, $\forall 1 \leq k < \min(n-1, p)$,

$$M_{k,\alpha} = \sup_{t \in \mathcal{T}_k} \left\{U_{k,t} - \overline{U^1_{k,t}}^{-1}(\alpha_{k,t})\right\}, \tag{12}$$

where $\{\alpha_{k,t},\ t \in \mathcal{T}_k\}$ is a collection of number in]0,1[chosen in accordance to the following procedure:

P3. For all $t \in \mathcal{T}_k$, $\alpha_{k,t} = \alpha_{k,n}$ where $\alpha_{k,n}$ is the α-quantile of the random variable

$$\inf_{t \in \mathcal{T}_k} \overline{U^1_{k,t}}\left\{U^1_{k,t}\right\}.$$

The null hypothesis $\hat{H}_k$ is rejected when $M_{k,\alpha}$ is positive. The calculation of the collection $\{\alpha_{k,t},\ t \in \mathcal{T}_k\}$ with the procedure P3 ensures that $\mathbb{P}\left((M_{k,\alpha} > 0) \cap A_k\right) \leq \alpha$.

In summary, the two-step procedure 'A' when σ is known is the following:

Procedure 'A'

1. Order the variables taking into account the observation Y,
2. (a) Set $\alpha \in (0,1)$,
   (b) For $1 \leq k < \min(n-1, p)$ calculate $M_{k,\alpha}$, defined by (12),
   (c) If it exists $1 \leq k < \min(n-1, p)$ such that $M_{k,\alpha}$ is non positive,
   Estimate the set of relevant variables J by $\hat{J} = \left\{(1), \ldots, (\mathring{k}_A)\right\}$ where $\mathring{k}_A = \inf\{k \geq 1, M_{k,\alpha} \leq 0\}$.
   Else $\hat{J} = \{(1), \ldots, (\min(n-1, p))\}$

The testing procedure 'A' is proved to be powerful and we give an upper bound of the probability to wrongly estimate J in the next theorem.

Theorem 3.2. *Let Y obey to Model (1). Assume that conditions A_1 and A_2 are verified. We denote by J the set $\{j, \beta_j \neq 0\}$ and by k_0 its cardinality. Let α and γ be fixed in $]0,1[$. The procedure 'A' estimates J by $\hat{J} = \left\{(1), \ldots, (\mathring{k}_A)\right\}$ where $\mathring{k}_A = inf\{k \geq 1, M_{k,\alpha} \leq 0\}$, where $M_{k,\alpha}$ is defined by (12) and $\{\alpha_{k,t}, t \in \mathcal{T}_k\}$ is defined according to the procedure P3.*

We consider the condition $(R_{2,k})$ for $k < k_0$ stated as

$(R_{2,k})$: $\exists t \leq \log_2(k_0 - k)$ *such that*

$$\frac{1}{2\sigma^2} inf\big\{||\Pi_S \mu||_n^2, S \in B_{2^t}\big\} \geq \frac{2^t}{n}\left[10 + 4\log\left(\frac{(p-k)k_0}{2^{2t}}\right)\right] + \frac{2}{n}\left[\sqrt{2^{t+1}\log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)} + \log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)\right],$$

where $\forall d \leq k_0, B_d = \{span(X_I), I \subset J, |I| = d\}$ *and* $|\mathcal{T}_k| = \lfloor \log_2(min(n-1,p) - k)\rfloor + 1$.

If $\forall k \leq k_0 - 1$ *the condition* $(R_{2,k})$ *holds, then*

$$\mathbb{P}_\mu(\hat{J} \neq J) \leq \gamma + \alpha + \delta, \tag{13}$$

where $\delta = \mathbb{P}_\mu(A_{k_0}^c) = P_\mu(\exists\ j \leq k_0/\beta_{(j)} = 0)$.

This theorem is non asymptotic and its result differs from Theorem 2.1 on the right part of (13). Indeed, the weight of the first step of the procedure, which lies in δ, was not involved in Section 2 since we considered ordered variable selection. Recall that Theorem 3.2 applies whatever the first step of the procedure. Moreover, the price to pay for a inadequate method chosen to order the variables appears clearly in (13) through δ. Indeed, Theorem 3.2 shows that the first step is essential in the two-step procedure 'A', which is easily understandable since there is no chance of having $J = \hat{J}$ if the event A_{k_0} does not occur at the end of the first step of procedure 'A'. Moreover, the condition $(R_{2,k})$ is also more restrictive than the condition (R_k) which appeared in Theorem 2.1.

Conditions on β_J explicitly appear in Theorem 3.2 when $\{X_i\}_{1\leq i\leq p}$ is an orthonormal family, see A.2.

3.2 Non-ordered variable selection with unknown variance

In this section, we define a procedure 'B' under the assumption that the variance σ^2 is unknown. Assume that the first step of the procedure 'B' has already been done: variables have been ordered.
In this section, the notations of Section 3.1 are used: $\forall\ 1 \leq k < \min(n-1,p)$, $t_{max}^k = \lfloor \log_2(\min(n-1,p) - k)\rfloor$, $\mathcal{T}_k = \big\{0, \ldots, t_{max}^k\big\}$.
We define $V_{(k)} = \text{span}(X_{(1)}, \ldots, X_{(k)})$ and $\forall\, t \in \mathcal{T}_k$, $S_{(k),(t)} = \text{span}\left(\Pi_{V_{(k)}^\perp}(X_{(k+1)}), \ldots, \Pi_{V_{(k)}^\perp}(X_{(k+2^t)})\right)$.
Denote for each $k \in \{1, \ldots, \min(n-1,p) - 1\}$, $t \in \mathcal{T}_k$, $V_{(k),(t)} = V_{(k)} \oplus S_{(k),(t)}$, and denote $D_{k,t} = 2^t$ and $N_{k,t} = n - (k + 2^t)$ the dimension of $S_{(k),(t)}$ and $V_{(k),(t)}^\perp$ respectively.

Since the variance is assumed to be unknown, we introduce for all $1 \leq k < \min(n-1,p)$ and for all $t \in \mathcal{T}_k$,

$$\tilde{U}_{D_{k,t},N_{k,t}} = \frac{N_{k,t}||\Pi_{S_{(k),(t)}} Y||_n^2}{D_{k,t}||Y - \Pi_{V_{(k),(t)}} Y||_n^2}.$$

In order to test the null hypothesis $\hat{H}_k$ defined by (10), we introduce a multiple testing procedure which relies this time on the statistics $\left\{\tilde{U}_{D_{k,t},N_{k,t}}, t \in \mathcal{T}_k\right\}$.

As in Section 3.1, we first provide a stochastic upper bound for the statistics $\tilde{U}_{D_{k,t},N_{k,t}}$ in order to define the multiple testing procedure.

Let $\epsilon' \sim \mathcal{N}_n(0, \sigma^2 I_n)$. Set for all $1 \leq k < \min(n-1, p)$ and for all $t \in \mathcal{T}_k$,

$$\Upsilon_{k,t} = \frac{N_{k,t}||\Pi_{S_{(k),\sigma_1^k(t)}}\epsilon'||_n^2}{D_{k,t}||\epsilon' - \Pi_{V_{(k),\sigma_1^k(t)}}\epsilon'||_n^2},$$

where $S_{(k),\sigma_1^k(t)} = \text{span}\left(\Pi_{V_{(k)}^\perp}(X_{\sigma_1^k(k+1)}), \ldots, \Pi_{V_{(k)}^\perp}(X_{\sigma_1^k(k+2^t)})\right)$, $V_{(k),\sigma_1^k(t)} = S_{(k),\sigma_1^k(t)} \oplus V_{(k)}$ and the permutation σ_1^k is defined as in Section 3.1.

Lemma 3.3. *Let $1 \leq k < min(n-1, p)$ and $t \in \mathcal{T}_k$.*
We define $A_k = \left\{\left\{X_{(1)}, \ldots, X_{(k)}\right\} = \{X_j, j \in J\}\right\}$. For all $x > 0$ we have

$$\mathbb{P}\left((\tilde{U}_{D_{k,t},N_{k,t}} > x) \cap A_k\right) \leq \mathbb{P}\left(\Upsilon_{k,t} > x\right).$$

Let $\bar{\Upsilon}_{k,t}(u)$ denote the probability for the statistic $\Upsilon_{k,t}$ to be larger than u.
Set $\forall \alpha \in]0,1[$, $\forall 1 \leq k < \min(n-1, p)$,

$$\hat{M}_{k,\alpha} = \sup_{t \in \mathcal{T}_k}\left\{\tilde{U}_{D_{k,t},N_{k,t}} - \bar{\Upsilon}_{k,t}^{-1}(\alpha_{k,t})\right\}, \tag{14}$$

where $\{\alpha_{k,t},\ t \in \mathcal{T}_k\}$ is a collection of number in]0,1[chosen in accordance to the following procedure:

P4. For all $t \in \mathcal{T}_k$, $\alpha_{k,t} = \alpha_{k,n}$ where $\alpha_{k,n}$ is the α-quantile of the random variable

$$\inf_{t \in \mathcal{T}_k} \bar{\Upsilon}_{k,t}\left\{\Upsilon_{k,t}\right\},$$

The null hypothesis $\hat{H}_k$ is rejected when $\hat{M}_{k,\alpha}$ is positive.The calculation of the collection $\{\alpha_{k,t},\ t \in \mathcal{T}_k\}$ with the procedure P4 ensures that $\mathbb{P}\left((\hat{M}_{k,\alpha} > 0) \cap A_k\right) \leq \alpha$.

In summary, the two-step procedure 'B' when σ is unknown is the following:

Procedure 'B'

1. Order the variables taking into account the observation Y,
2. (a) Set $\alpha \in (0,1)$,
   (b) For $1 \leq k < \min(n-1, p)$ calculate $\hat{M}_{k,\alpha}$, defined by (14),
   (c) If it exists $1 \leq k < \min(n-1, p)$ such that $\hat{M}_{k,\alpha}$ is non positive,
   Estimate the set of relevant variables J by $\hat{J} = \left\{(1), \ldots, (\mathring{k}_B)\right\}$
   where $\mathring{k}_B = \inf\left\{k \geq 1, \hat{M}_{k,\alpha} \leq 0\right\}$.
   Else $\hat{J} = \{(1), \ldots, (\min(n-1, p))\}$

The procedure 'B' is proved to be powerful in the next theorem; we give an upper bound of the probability to wrongly estimate J under some conditions on the signal. Let us introduce some notations that will be used in the following theorem. We set
$L_t = \log(|\mathcal{T}_k|/\alpha)$, $m_t = \exp(4L_t/N_{k,t})$, $m_p = \exp\left(\frac{4D_{k,t}}{N_{k,t}}\log\left(\frac{e(p-k)}{D_{k,t}}\right)\right)$, $M = 2m_t m_p$.
Denote $\Lambda_1(k,t) = \sqrt{1+\frac{D_{k,t}}{N_{k,t}}}$, $\Lambda_2(k,t) = \left(1+2\frac{D_{k,t}}{N_{k,t}}\right)M$ and $\Lambda_3(k,t) = 2\Lambda_1(k,t) + \Lambda_2(k,t)$.

Theorem 3.4. *Let Y obey to model (1). Assume that conditions A_1 and A_2 are verified. We define by J the set $\{j, \beta_j \neq 0\}$ and by k_0 its cardinality. Let α and γ be fixed in $]0,1[$. The procedure 'B' estimates J by $\hat{J} = \left\{(1), \ldots, (\mathring{k}_B)\right\}$ where $\mathring{k}_B = inf\left\{k \geq 1, \hat{M}_{k,\alpha} \leq 0\right\}$, where $\hat{M}_{k,\alpha}$ is defined by (14) and $\{\alpha_{k,t}, t \in \mathcal{T}_k\}$ is defined according to the procedure P4.*

We consider the condition $(R_{3,k})$ for $k < k_0$ stated as
$(R_{3,k})$: *$\exists t \leq \log_2(k_0 - k)$ such that*

$$\begin{aligned}\frac{1}{2}inf\left\{||\Pi_S\mu||_n^2, S \in B_{2^t}\right\} &\geq \frac{A(k,t)}{N_{k,t}}\left[||\mu||_n^2 + \sigma^2\left(2+\frac{3}{n}\log\left(\frac{2k_0}{\gamma}\right)\right)\right] \\ &+ \frac{\sigma^2}{n}\left[2^t\left(6+4\log\left(\frac{k_0}{2^t}\right)\right)+3\log\left(\frac{2k_0}{\gamma}\right)\right],\end{aligned}$$

where

$$\begin{aligned}A(k,t) &= 2^t\left[2+\frac{2^t}{N_{k,t}}+\Lambda_3(k,t)\log\left(\frac{e(p-k)}{2^t}\right)\right] \\ &+ (1+\Lambda_2(k,t))\log\left(\frac{\log_2(p-k)+1}{\alpha}\right),\end{aligned}$$

and $\forall d \leq k_0, B_d = \{span(X_I), I \subset J, |I| = d\}$.

If $\forall k \leq k_0 - 1$ the condition $(R_{3,k})$ holds, then

$$\mathbb{P}_\mu(\hat{J} \neq J) \leq \gamma + \alpha + \delta, \tag{15}$$

where $\delta = \mathbb{P}_\mu(A_{k_0}^c) = P_\mu(\exists\ j \leq k_0/\beta_{(j)} = 0)$.

This theorem is non asymptotic and shows that the testing procedure 'B' is powerful under some conditions on the signal. As for Theorem 3.2 of Section 3.1, the first step of the procedure -the ordering of the variables- has an important part in Theorem 3.4. A simulation study in Section 4 will show that this testing procedure combined with a good way to order variables -in order to minimize δ- performs well.

Remark 3.5. *The condition $(R_{3,k})$ can be simplified under the assumption that $2^t \leq (n-k)/2$ and $\log(p-k) > 1$. Indeed, in this case, the right hand in condition $(R_{3,k})$ is upper*

bounded by

$$C(||\mu||_n, \gamma, \alpha, \sigma) 2^t \left[\frac{\log(p-k)}{N_{k,t}} + \frac{\log(k_0)}{n} \right], \tag{16}$$

where $C(||\mu||_n, \gamma, \alpha, \sigma)$ *is a constant depending on* $||\mu||_n$, γ, α *and* σ.

Conditions on β_J explicitly appear in Theorem 3.4 when $\{X_i\}_{1\leq i\leq p}$ is an orthonormal family, see A.3.

4 Simulation study

4.1 Presentation of the procedures

In this section, we comment the results of the simulation study which are presented in tables 1-4. Our aim was to compare the performances of our selection methods. The procedures presented in this paper are implemented in the R-package *mht* which is available on CRAN (http://cran.r-project.org/). Six methods were compared; the procedure described in Section 2 for ordered variable selection, denoted "proc-ordered" in the tables, the two-step procedure 'B' described in Section 3, either with ordered p-values denoted "procpval" or with the Bolasso order denoted "procbol", the FDR procedure described in Bunea et al. (2006), the Lasso method and the Bolasso technique. The comparison of the first method and the others is unfair and was not performed because the information of the relative importance of the variables is known for ordered variable selection. The two kinds of method have to be separately compared.

The simulation was performed when $(X_i)_{1\leq i\leq p}$ is a linearly independent family and in the high-dimensional case ($p \geq n$). For the latter, the FDR procedure of Bunea et al. (2006) cannot be computed as p-values cannot be obtained with the least squares estimate with all p variables. In this case we compared an adjusted FDR; a p-value was calculated for each variable X_i from the regression of Y onto that variable. As mentioned in the introduction, this is a natural extension of the FDR procedure in high-dimensional analysis and extended FDR is widely used in biology for differential and transcriptome analysis.

When $(X_i)_{2\leq i\leq p}$ is a linearly independent family, the calculation of $T_{k,\alpha}$ with (3) -for ordered variable selection- requires a high computational time, as a calculation of $V_k^\perp$ and $\{S_{k,t}, t \in \mathcal{T}_k\}$ is needed for each k. Since a variable selection method is not only judged on its results but also on its fastness, useless calculations in our procedure had to be avoided. The Gram-Schmidt process was used to get an orthonormal family out of $(X_i)_{2\leq i\leq p}$. Thus the calculation of $(V_k^\perp)_{k\geq 0}$ was done once and for all.
Decompose $\forall\ l > 0$: $X_{k+l} = \Pi_{V_k}(X_{k+l}) + \Pi_{V_k^\perp}(X_{k+l})$. Note $(e_j)_{j=1,\dots,k}$ an orthonormal basis of V_k, then:
$\Pi_{V_k}(X_{k+l}) = \sum_{j=1}^k < X_{k+l}, e_j > e_j$ and $\Pi_{V_k^\perp}(X_{k+l}) = X_{k+l} - \sum_{j=1}^k < X_{k+l}, e_j > e_j$. The family $(X_1, \dots, X_p)$ was modified into

$$\left(X_1, \frac{\Pi_{V_1^\perp}(X_2)}{||\Pi_{V_1^\perp}(X_2)||_n}, \frac{\Pi_{V_2^\perp}(X_3)}{||\Pi_{V_2^\perp}(X_3)||_n}, \dots, \frac{\Pi_{V_{p-1}^\perp}(X_p)}{||\Pi_{V_{p-1}^\perp}(X_p)||_n} \right).$$

Denote that orthonormal family by $(\tilde{X}_1, \ldots, \tilde{X}_p)$. We decomposed Y as:

$$Y = \underbrace{\tilde{X}_1\tilde{\beta}_1 + \cdots + \tilde{X}_k\tilde{\beta}_k}_{V_k} + \underbrace{\tilde{X}_{k+1}\tilde{\beta}_{k+1} + \cdots + \tilde{X}_p\tilde{\beta}_p}_{\subset V_k^{\perp}} + \epsilon. \tag{17}$$

Then $S_{k,t} = \text{span}(\tilde{X}_{k+1}, \ldots, \tilde{X}_{k+2^t})$ and so $\left|\left|\Pi_{S_{k,t}} Y\right|\right|_n^2 = \tilde{\beta}_{k+1}^2 + \cdots + \tilde{\beta}_{k+2^t}^2$. This technique avoided a lot of useless and redundant calculations.
The decomposition of Gram-Schmidt has also been used in the non-ordered variable selection case with the two-step procedure 'A' and 'B' once the variables have been ordered.

4.2 Design of our simulation study

Concerning the design of our simulations, we set X_1 to be the vector of $\mathbb{R}^n$ whose coordinates are all equal to 1 and we considered four models. For each model we computed $Y = \beta_{j_1} X_{j_1} + \cdots + \beta_{j_{k_0-1}} X_{j_{k_0-1}} + \epsilon$, where ϵ is a vector of independent standard Gaussian variables, $J = \{1, j_1, \ldots, j_{k_0-1}\} \subset \{1, \ldots, p\}$. Models differ in how the response variable Y is linked to the X_i and the dependence structure of the X_i's. The models are defined as follows:

(A) Simple model: we simulated $p-1$ independent vectors $X_i^* \sim \mathcal{N}_n(0, I_n)$ and β_J are all equal to $10/\sqrt{n}$ or $6/\sqrt{n}$.

(B) Correlated model: $X_2^* \sim \mathcal{N}_n(0, I_n)$ and for $i \geq 2, X_{i+1}^* = \rho X_i^* + (1-\rho^2)^{1/2}\epsilon_i^*$, where $\rho = 0.5$, $(\epsilon_i^*, i = 2, \ldots, p)$ are independent vectors $\epsilon_i^* \sim \mathcal{N}_n(0, I_n)$ and β_J are all equal to $10/\sqrt{n}$ or $6/\sqrt{n}$.

(C) Triangle model: $\beta_j = \gamma(11-j), j = 2, \ldots, 11, \beta_j = 0, j > 11$ and we simulated $p-1$ independent vectors $X_i^* \sim \mathcal{N}_n(0, I_n)$.

(D) Correlated Triangle model: as C, but with $X_{i+1}^* = \rho X_i^* + (1-\rho^2)^{1/2}\epsilon_i^*$, where $\rho = 0.5$, $(\epsilon_i^*, i = 2, \ldots, p)$ are independent vectors $\epsilon_i^* \sim \mathcal{N}_n(0, I_n)$.

In all models, we set the predictors $X_i = X_i^*/||X_i^*||_n^2$, for $i = 2, \ldots, p$. Thus $\frac{1}{n}\sum_{j=1}^n X_{ij}^2 = 1$. We considered two instances of k_0 for model A and B (6 or 11). Models B and C have been simulated with $\gamma = 0.5$ in a low dimensional setting and $\gamma = 1$ in a high dimensional context. Note that the choice of γ in a low dimensional setting gives the same signal as the simulation in Wasserman and Roeder (2009) but a lower signal in the high dimensional setting as they used $\gamma = 1.5$. For models A and B, samples of $n = 100$ and $p = 80$ or 600 have been simulated; $n = 100$ and $p = 80$ or 1000 for models C and D. In each case, 500 replications have been made.

In all models, we were interested in the percentage of true model recovered (labelled as "Truth"), which is the number of time we actually found $\hat{J} = J$ over the total number of simulations. We also recorded the number of selected variables ("Inclusions"), the number of relevant variables that were included in the selected model ("Correct incl.") and the MSE (Mean Squared Error) which was calculated by average over all simulations:

$MSE = \sum_{j=1}^{n}(\hat{Y}_j - (X\beta_J)_j)/n$, where $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ is an estimation of β by linear regression on the selected set of variables $(X_j, j \in \hat{J})$.

Since the first step of our procedure is an important step, we were also interested in an estimation of the probability to obtain a wrong order on the variables, depending on the method that has been used ($\delta = P_\mu(A_{k_0}^c)$). This estimation is not mentioned for the ordered variable selection procedure.

The FDR procedure described in Bunea et al. (2006) was set by choosing q (user level) as 0.1 and 0.05.

The l^1 penalty of the Lasso was tuned via 10-cross validation. Concerning the Bolasso technique, we choose $it = 100$ bootstrap iterations; the penalty was also tuned via 10-cross validation. Both methods were always ended with a linear regression on the estimated set of indices of the relevant variables in order to minimize the bias of the Lasso method.

When the Bolasso technique was used to order the variable at the first step of procedure 'B', we chose to stop the dichotomy algorithm (see Section 3) as soon as 60 variables were ordered. The objective was to spare calculation since it was uneasy to distinguish the remaining variables after the sixtieth position.

Concerning the three procedures presented in this paper, the results are displayed for a level $\alpha \in \{0.1, 0.05\}$. For ordered variable selection, $(X_i)_{1\leq i\leq p}$ was modified into $(X_J, X_{\{1,\dots,p\}\setminus J})$ and the collection $\{\alpha_{k,t}, t \in \mathcal{T}_k\}$ was chosen in accordance to the procedure P1, which required more computational time than P2, but which was far much powerful. For non-ordered variable selection, the collection $\{\alpha_{k,t}, t \in \mathcal{T}_k\}$ was chosen with the procedure P4, since the variance was considered unknown in the simulation.

4.3 Comments on the results

Table 1 through Table 4 present the results of model A through model D, respectively.

In all tables, the procedure of multiple hypotheses testing developed for ordered variable selection in Section 2 gave excellent results, even in the high-dimensional case where $p > n$. These results are not surprising because our choices of β ensured that at each step the tests are powerful, so the probability of wrongly estimating k_0 was almost reduced to α.

The method "procbol" is shown to give the best results over the tested method for non-ordered variable selection. Indeed, the percentage of true model recovered is the highest and the MSE is the lowest among the tested methods, even in the correlated models B and D. Note that the difference between our signal and the one of Wasserman and Roeder (2009) in the high-dimensional context ($\gamma = 1$ against $\gamma = 1.5$) was motivated by the results of the "procbol" method when we set $\gamma = 1.5$: the results were close to perfection for both our method and the Lasso, thus we decided to simulate a lower signal in order to show differences in the results. This shows that splitting the data is not essential to obtain good results, as observed in Wasserman and Roeder (2009). However, a combination of a small β_J and a high number of variables induced a high $\hat{\delta}$ and consequently decreased the power of our "procbol" method. Moreover, the results of the "procbol" method become less satisfactory -but still better than the other methods on our simulations- with an increase on the value of k_0 because of the overestimation of the statistics in Lemma 3.3.

The simulations confirmed that the first step of our procedure, namely the ordering

of the variables, is a crucial step. Indeed, the difference of results between "procpval" and "procbol" was striking and only relied on the order which was given to the variables. Thus the Bolasso technique is to be preferred for ordering the variables, at least in our simulations.

Since the FDR method is based on the same order as "procpval" -the p-values-, the difference between the two methods lies in where the cut-off between the relevant variables and the others is. On that matter, the "procpval" method gave better results in that the MSE is lower and the number of correct variables included is higher, showing that the multiple testing procedure presented in Section 3 improved the estimation of the set of relevant variables over the threshold used by the FDR procedure.

Results	proc-ordered		procpval		procbol		FDR		Lasso	Bolasso
	α=0.1	α=0.05	α=0.1	α=0.05	α=0.1	α=0.05	q=0.1	q=0.05		
$k_0 = 11, n = 100, p = 80, \beta_J = 10/\sqrt{n}$										
			$\hat{\delta} = 0.46$		$\hat{\delta} = 0.00$		$\hat{\delta} = 0.45$			
Truth	0.92	0.96	0.54	0.54	0.94	0.96	0.13	0.10	0.29	0.67
Inclusions	11.33	11.15	13.06	12.62	11.08	11.05	8.55	7.60	13.18	11.70
Correct incl.	11.00	11.00	10.92	10.90	11.00	11.00	8.34	7.53	11.00	10.99
MSE	0.12	0.11	0.20	0.22	0.11	0.11	2.97	3.72	0.18	0.14
$k_0 = 6, n = 100, p = 80, \beta_J = 6/\sqrt{n}$										
			$\hat{\delta} = 0.88$		$\hat{\delta} = 0.07$		$\hat{\delta} = 0.82$			
Truth	0.91	0.95	0.11	0.11	0.86	0.84	0.00	0.00	0.27	0.47
Inclusions	6.37	6.13	7.30	6.54	6.00	5.94	1.98	1.66	8.22	7.14
Correct incl.	6.00	6.00	5.05	4.90	5.91	5.87	1.86	1.62	5.94	5.94
MSE	0.06	0.06	0.40	0.44	0.08	0.09	1.42	1.45	0.16	0.13
$k_0 = 11, n = 100, p = 600, \beta_J = 10/\sqrt{n}$										
			$\hat{\delta} = 1.00$		$\hat{\delta} = 0.17$		$\hat{\delta} = 1.00$			
Truth	0.89	0.95	0.00	0.00	0.83	0.83	0.00	0.00	0.00	0.25
Inclusions	11.66	11.21	5.88	5.36	11.30	11.20	3.33	3.02	17.97	13.24
Correct incl.	11.00	11.00	5.68	5.23	10.99	10.99	3.33	3.02	10.99	10.99
MSE	0.12	0.11	4.11	4.56	0.11	0.11	6.34	6.69	0.31	0.20
$k_0 = 6, n = 100, p = 600, \beta_J = 6/\sqrt{n}$										
			$\hat{\delta} = 0.95$		$\hat{\delta} = 0.30$		$\hat{\delta} = 0.92$			
Truth	0.91	0.96	0.05	0.05	0.62	0.56	0.00	0.00	0.11	0.26
Inclusions	6.43	6.12	4.36	4.22	5.62	5.48	2.48	2.18	11.52	8.49
Correct incl.	6.00	6.00	4.14	4.04	5.50	5.39	2.46	2.17	5.59	5.65
MSE	0.06	0.06	0.59	0.62	0.22	0.25	1.10	1.22	0.37	0.30

Table 1: Model A. Results of the 500 simulations. The first row gives an estimation of $\delta = P_\mu(A_{k_0}^c)$. The second row "Truth" records the pourcentage of time $\hat{J} = J$. "Inclusions" records the number of selected variables and "Correct incl." the number of selected variables that are relevant. The MSE is calculated by average over all simulations.

Results	proc-ordered		procpval		procbol		FDR		Lasso	Bolasso
	α=0.1	α=0.05	α=0.1	α=0.05	α=0.1	α=0.05	q=0.1	q=0.05		
$k_0 = 11, n = 100, p = 80, \beta_J = 10/\sqrt{n}$										
			$\hat{\delta} = 0.$		$\hat{\delta} = 0.02$		$\hat{\delta} = 0.91$			
Truth	0.90	0.96	0.09	0.09	0.94	0.96	0.00	0.00	0.36	0.83
Inclusions	11.36	11.14	15.12	14.52	11.08	11.05	4.63	3.65	12.52	11.28
Correct incl.	11.00	11.00	10.35	10.28	11.00	11.00	4.44	3.55	11.00	11.00
MSE	0.14	0.13	0.58	0.63	0.12	0.12	9.80	11.6	0.16	0.12
$k_0 = 6, n = 100, p = 80, \beta_J = 6/\sqrt{n}$										
			$\hat{\delta} = 0.99$		$\hat{\delta} = 0.07$		$\hat{\delta} = 0.99$			
Truth	0.90	0.95	0.01	0.01	0.76	0.74	0.00	0.00	0.60	0.64
Inclusions	6.45	6.20	8.27	7.45	5.96	5.84	1.44	1.28	7.03	6.21
Correct incl.	6.00	6.00	4.30	4.17	5.83	5.77	1.33	1.23	5.98	5.77
MSE	0.08	0.07	0.65	0.70	0.11	0.12	2.49	2.57	0.11	0.15
$k_0 = 11, n = 100, p = 600, \beta_J = 10/\sqrt{n}$										
			$\hat{\delta} = 1.00$		$\hat{\delta} = 0.09$		$\hat{\delta} = 1.00$			
Truth	0.91	0.96	0.00	0.00	0.90	0.91	0.00	0.00	0.00	0.58
Inclusions	11.41	11.15	3.75	3.19	11.11	11.10	2.78	2.20	16.54	11.63
Correct incl.	11.00	11.00	3.64	3.12	11.00	11.00	2.78	2.19	10.99	10.99
MSE	0.14	0.13	6.03	6.57	0.12	0.12	6.86	7.63	0.29	0.16
$k_0 = 6, n = 100, p = 600, \beta_J = 6/\sqrt{n}$										
			$\hat{\delta} = 0.84$		$\hat{\delta} = 0.13$		$\hat{\delta} = 0.84$			
Truth	0.89	0.95	0.15	0.15	0.73	0.71	0.03	0.01	0.36	0.51
Inclusions	6.44	6.14	5.20	5.02	5.82	5.72	4.68	3.93	8.22	7.04
Correct incl.	6.00	6.00	4.85	4.78	5.74	5.67	4.15	3.62	5.88	5.84
MSE	0.07	0.07	0.45	0.49	0.15	0.17	0.69	0.88	0.19	0.19

Table 2: Model B. Results of the 500 simulations. The first row gives an estimation of $\delta = P_\mu(A_{k_0}^c)$. The second row "Truth" records the pourcentage of time $\hat{J} = J$. "Inclusions" records the number of selected variables and "Correct incl." the number of selected variables that are relevant. The MSE is calculated by average over all simulations.

Results	proc-ordered		procpval		procbol		FDR		Lasso	Bolasso
	α=0.1	α=0.05	α=0.1	α=0.05	α=0.1	α=0.05	q=0.1	q=0.05		
$n = 100, p = 80$										
			$\hat{\delta} = 0.71$		$\hat{\delta} = 0.11$		$\hat{\delta} = 0.71$			
Truth	0.85	0.91	0.26	0.25	0.80	0.75	0.05	0.04	0.23	0.50
Inclusions	10.47	10.20	9.61	9.48	9.94	9.83	9.15	8.96	11.65	10.45
Correct incl.	9.99	9.98	9.37	9.33	9.86	9.79	8.99	8.89	9.77	9.81
MSE	0.12	0.11	0.24	0.25	0.14	0.15	0.41	0.50	0.23	0.18
$n = 100, p = 1000$										
			$\hat{\delta} = 1.00$		$\hat{\delta} = 0.05$		$\hat{\delta} = 1.00$			
Truth	0.93	0.96	0.00	0.00	0.95	0.95	0.00	0.00	0.03	0.76
Inclusions	10.38	10.14	5.04	5.04	10.06	10.05	3.18	3.00	13.38	10.29
Correct incl.	10.00	10.00	5.00	5.00	10.00	10.00	3.18	3.00	10.00	10.00
MSE	0.11	0.10	60.97	60.97	0.10	0.11	118	124	0.18	0.12

Table 3: Model C. Results of the 500 simulations. The first row gives an estimation of $\delta = P_\mu(A^c_{k_0})$. The second row "Truth" records the pourcentage of time $\hat{J} = J$. "Inclusions" records the number of selected variables and "Correct incl." the number of selected variables that are relevant. The MSE is calculated by average over all simulations.

Results	proc-ordered		procpval		procbol		FDR		Lasso	Bolasso
	α=0.1	α=0.05	α=0.1	α=0.05	α=0.1	α=0.05	q=0.1	q=0.05		
$n = 100, p = 80$										
			$\hat{\delta} = 0.49$		$\hat{\delta} = 0.09$		$\hat{\delta} = 0.49$			
Truth	0.90	0.94	0.44	0.43	0.71	0.67	0.08	0.05	0.45	0.60
Inclusions	10.31	10.08	10.07	10.00	9.86	9.73	8.55	8.12	10.92	10.29
Correct incl.	9.99	9.98	9.58	9.54	9.76	9.69	8.40	8.06	9.82	9.84
MSE	0.11	0.11	0.22	0.22	0.15	0.16	1.48	2.20	0.18	0.15
$n = 100, p = 1000$										
			$\hat{\delta} = 1.00$		$\hat{\delta} = 0.00$		$\hat{\delta} = 1.00$			
Truth	0.87	0.93	0.00	0.00	0.97	0.99	0.00	0.00	0.85	0.85
Inclusions	10.65	10.31	13.05	13.03	10.03	10.01	7.00	6.82	10.52	10.75
Correct incl.	10.00	10.00	8.99	8.99	10.00	10.00	7.00	6.82	10.00	10.00
MSE	0.12	0.11	0.78	0.79	0.11	0.10	16.7	22.9	0.13	0.14

Table 4: Model D. Results of the 500 simulations. The first row gives an estimation of $\delta = P_\mu(A^c_{k_0})$. The second row "Truth" records the pourcentage of time $\hat{J} = J$. "Inclusions" records the number of selected variables and "Correct incl." the number of selected variables that are relevant. The MSE is calculated by average over all simulations.

5 Conclusion

This paper tackled the problem of recovering the set of relevant variables J in a sparse linear model, especially when the number of variables p was higher than the sample size n. We proposed new methods based on hypotheses testing to estimate J. The procedure presented for non-ordered variables selection with unknown variance is a two-step procedure that needs to be combined with a good method to order the variables (first step). The procedure applies with any possible order; we propose the use of the Bolasso technique and it should be preferred to an order obtained from p-values (as the FDR procedure) as it gave better results on simulations. The procedures are proved to be powerful under some conditions on the signal and the theorems are non asymptotic. The simulations showed that these new procedures outperformed all the other tested methods, especially in the high-dimensional case, which was the aim of this study.

Acknowledgements

The author is grateful to Beatrice Laurent and Magali San-Cristobal for helpful comments and suggestions.

References

Bach, F. (2009). Model-consistent sparse estimation through the bootstrap.

Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptative test of linear hypotheses by model selection. *Ann. Statist.*, 31(1):225–251.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *J. R. Stat. Soc.*, B 57, 289-300.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electron. J. Statist.*, 1:169–194.

Bunea, F., Wegkamp, M., and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Statist. Plann. Inference*, 136:4349–4363.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Chesneau, C. and Hebiri, M. (2008). Some theoretical results on the grouped variables lasso. *Math. Methods Statist.*, 17(4):317–326.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptative lasso for sparse high-dimensional regression models. *Stat. Sin.*, 18(4):1603–1618.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338.

Lê Cao, K. A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7:Article 35.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Tenenhaus, M. (1998). *La régression PLS: théorie et pratique*. Editions Technip.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.*, B 58(1):267–288.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ^1-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on*, 55:2183–2202.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.*, 37:2178–2201.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563.

A Variables selection when $\{X_i\}_{1\leq i\leq p}$ is an orthonormal family

A.1 Ordered variable selection when $\{X_i\}_{1\leq i\leq p}$ is an orthonormal family

When $(X_i)_{2\leq i\leq p}$ is an orthonormal family and the variance is unknown, an other upper bound of the statistics $\tilde{U}_{D_{k,t},N_{k,t}}$ than the one in Lemma 3.3 can be used. Indeed, we can obtain an upper bound which does not depend on the family $(X_i)_{1\leq i\leq p}$ nor on the order on that family.
Let $I_1,\dots,I_p$ be p i.i.d. standard Gaussian variables, and let $|I_{(1)}|\geq\cdots\geq|I_{(p)}|$.
We define: $\forall k=0,\dots,p-1,\ \forall D=0,\dots,p-k-1,\ L_{k,D}=\sum_{j=k+D+1}^{p} I_{(j)}^2$.
Let $1\leq k<p$ and $t\in\mathcal{T}_k$, we define
$A_k=\left\{\left\{X_{(1)},\dots,X_{(k)}\right\}=\{X_j,j\in J\}\right\}$. For all $x>0$ we have

$$\mathbb{P}\left((\tilde{U}_{D_{k,t},N_{k,t}}>x)\cap A_k\right)\leq\mathbb{P}\left(\frac{N_{k,t}}{D_{k,t}}\frac{Z_{D_{k,t},p-k}}{L_{k,D_{k,t}}+K_{n-p}}>x\right),$$

where K_{n-p} is a chi-square variable with $n-p$ degrees of freedom and $Z_{d,D}$ is defined by (18).

A.2 Non ordered variables selection when $\{X_i\}_{1\leq i\leq p}$ is an orthonormal family and the variance is known

When the family $(X_i)_{2\leq i\leq p}$ is orthonormal, $U_{k,t}$ can be stochastically upper bounded by a statistic that does not depend on $(X_i)_{1\leq i\leq p}$ nor on the first step of the procedure.
Let $D>0$ and $W_1,\dots,W_D$ be D i.i.d. standard Gaussian variables ordered as $|W_{(1)}|\geq\cdots\geq|W_{(D)}|$.
We define: $\forall d=1,\dots,D$,

$$Z_{d,D}=\sum_{j=1}^{d}W_{(j)}^2. \tag{18}$$

Let $\bar{Z}_{d,D}(u)$ denote the probability for the statistic $Z_{d,D}$ to be larger than u.
A multiple testing procedure can be derived from procedure 'A' with this upper bound.

Lemma A.1. *Let $1\leq k<p$ and $t\in\mathcal{T}_k$.*
We define $A_k=\left\{\left\{X_{(1)},\dots,X_{(k)}\right\}=\{X_j,j\in J\}\right\}$. For all $x>0$, we have

$$\mathbb{P}\left((U_{k,t}>x)\cap A_k\right)\leq\mathbb{P}\left(Z_{D_{k,t},p-k}/n>x\right).$$

Set $\forall\alpha\in]0,1[$, $\forall 1\leq k<p$,

$$M_{k,\alpha}^1=\sup_{t\in\mathcal{T}_k}\left\{U_{k,t}-\bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_{k,t})/n\right\}, \tag{19}$$

where $\{\alpha_{k,t},\ t\in\mathcal{T}_k\}$ is a collection of number in]0,1[chosen in accordance to the following procedure:

P4. For all $t\in\mathcal{T}_k,\alpha_{k,t}=\alpha_{k,n}$ where $\alpha_{k,n}$ is the α-quantile of the random variable

$$\inf_{t\in\mathcal{T}_k}\bar{Z}_{D_{k,t},p-k}\left\{Z_{D_{k,t},p-k}\right\}.$$

The null hypothesis $\hat{H}_k$ is rejected when $M_{k,\alpha}^1$ is positive. The major benefit of Procedure 'A' when the family $(X_i)_{2\leq i\leq p}$ is orthonormal is that the upper bound of the statistics $U_{k,t}$ in Lemma A.1 does not depend on the family $(X_i)_{1\leq i\leq p}$ nor on the order on that family. Thus the collection $\{\alpha_{k,t},t\in\mathcal{T}_k\}$ defined by the procedure P4 only depends on k and t, with p and n fixed.

In the particular case where $(X_i)_{2\leq i\leq p}$ is an orthonormal family, we obtain the following corollary of Theorem 3.2, which is more explicit.

Corollary A.2. *Let Y obey to model (1). We assume that $p<n$ and that $(X_i)_{2\leq i\leq p}$ is an orthonormal family. We denote by J the set $\{j,\beta_j\neq 0\}$ and by k_0 its cardinality. Let α and γ be fixed in $]0,1[$.*
The procedure estimates J by $\hat{J}=\left\{(1),\dots,(\mathring{k}_{Abis})\right\}$ where $\mathring{k}_{Abis}=inf\left\{k\geq 1,M_{k,\alpha}^1\leq 0\right\}$, where $M_{k,\alpha}^1$ is defined by (19) and $\{\alpha_{k,t},t\in\mathcal{T}_k\}$ is defined according to the procedure P4.

We consider the condition $(R_{2bis,k})$ *for* $k < k_0$ *stated as*
$(R_{2bis,k}) : \exists\ t \leq \log_2(k_0 - k)$ *such that*

$$\begin{aligned}\frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta^2_{\sigma_2(j)} \;&\geq\; \frac{2^t}{n}\left[10 + 4\log\left(\frac{(p-k)k_0}{2^{2t}}\right)\right] \\ &\quad + \frac{2}{n}\left[\sqrt{2^{t+1}\log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)} + \log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)\right],\end{aligned}$$

where σ_2 *is defined by* $|\beta_{\sigma_2(1)}| \leq \cdots \leq |\beta_{\sigma_2(k_0)}|$ *and* $|\mathcal{T}_k| = \lfloor \log_2(p-k) \rfloor + 1$.

If $\forall k \leq k_0 - 1$ *the condition* $(R_{2bis,k})$ *holds, then*

$$\mathbb{P}_\mu(\hat{J} \neq J) \leq \gamma + \alpha + \delta, \tag{20}$$

where $\delta = \mathbb{P}_\mu(A^c_{k_0}) = P_\mu(\exists\ j \leq k_0 / \beta_{(j)} = 0)$.

A.3 Non ordered variables selection when $\{X_i\}_{1\leq i\leq p}$ is an orthonormal family and the variance is unknown

When $(X_i)_{2\leq i\leq p}$ is an orthonormal family, the condition $(R_{3,k})$ of Theorem 3.4 can be rewritten in a more explicit way. The new condition $(R_{3bis,k})$ obtained in this case is the following:
$(R_{3bis,k})$: $\exists t \leq \log_2(k_0 - k)$ such that

$$\begin{aligned}\frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta^2_{\sigma_2(j)} \;&\geq\; \frac{A(k,t)}{N_{k,t}}\left[\sum_{j=k+2^t}^{j=k_0}\beta^2_{\sigma_2(j)} + \sigma^2\left(2 + \frac{3}{n}\log\left(\frac{2k_0}{\gamma}\right)\right)\right] \\ &+\; \frac{\sigma^2}{n}\left[2^t\left(6 + 4\log\left(\frac{k_0}{2^t}\right)\right) + 3\log\left(\frac{2k_0}{\gamma}\right)\right],\end{aligned}$$

where σ_2 is defined such that $|\beta_{\sigma_2(1)}| \leq \cdots \leq |\beta_{\sigma_2(k_0)}|$ and $A(k,t)$ is defined as in Theorem 3.4.
Remark 3.5 is also verified in the particular case where $(X_i)_{2\leq i\leq p}$ is an orthonormal family.

The differences between the two conditions $(R_{3,k})$ and $(R_{3bis,k})$ lie in the fact that $\inf\{||\Pi_S\mu||^2_n, S \in B_{2^t}\} = \sum_{j=1}^{2^t}\beta^2_{\sigma_2(j)}$ and that the upper bound of $Q_{1-\gamma/2k_0}$ is modified, where $Q_{1-\gamma/2k_0}$ is defined by $\mathbb{P}\left(||Y - \Pi_{V_{(k),(t)}}Y||^2_n > Q_{1-\gamma/2k_0} \cap A_{k_0}\right) \leq \gamma/2k_0$.
Indeed, on the event A_{k_0} we have that $||Y - \Pi_{V_{(k),(t)}}Y||^2_n \leq \sum_{j=k+2^t}^{j=k_0}\beta^2_{\sigma_2(j)} + ||\epsilon||^2_n$, where σ_2 is defined such that $|\beta_{\sigma_2(1)}| \leq ... \leq |\beta_{\sigma_2(k_0)}|$. We get from there the condition $(R_{3bis,k})$.

B Proofs

Proof of Theorem 2.1. Let $k \leq k_0 - 1$ and assume that (R_k) holds. According to Baraud et al. (2003), the power of the test H_k, $\mathbb{P}_\mu(T_{k,\alpha} > 0)$, is greater than $1 - \gamma/k_0$. This

is equivalent to
$\mathbb{P}_\mu(H_k \text{ is accepted}) \leq \gamma/k_0$.
Moreover, for all $k_0 \leq k \leq \min(n-1, p)$, $\mathbb{P}_\mu(T_{k,\alpha} > 0) \leq \alpha$, since α is the level of the test H_k. Then we have:

$$\begin{aligned}\mathbb{P}_\mu(\hat{k} > k_0) &\leq \mathbb{P}_\mu(H_{k_0} \text{ is rejected}) = \mathbb{P}_\mu(T_{k_0,\alpha} > 0) \\ &\leq \alpha\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}_\mu(\hat{k} < k_0) &\leq \sum_{j=0}^{k_0-1} \mathbb{P}_\mu(H_j \text{ is accepted}) \\ &\leq k_0\gamma/k_0.\end{aligned}$$

Hence we obtain

$$\mathbb{P}_\mu(\hat{k} \neq k_0) \leq \mathbb{P}_\mu(\hat{k} < k_0) + \mathbb{P}_\mu(\hat{k} > k_0) \leq \gamma + \alpha,$$

which concludes the proof of (7).

□

Proof of Lemma 3.1. Let $x > 0$. By definition of $U_{k,t}$, we have
$\mathbb{P}\left((U_{k,t} > x) \cap A_k\right)$

$$\begin{aligned}&= \mathbb{P}\left(\left(\frac{||\Pi_{S_{(k),(t)}} Y||_n^2}{\sigma^2} > x\right) \cap A_k\right) \\ &= \mathbb{P}\left(\left(\frac{||\Pi_{S_{(k),(t)}} \mu||_n^2 + ||\Pi_{S_{(k),(t)}} \epsilon||_n^2}{\sigma^2} > x\right) \cap A_k\right).\end{aligned}$$

Since $A_k = \left\{\left\{X_{(1)}, \ldots, X_{(k)}\right\} = \{X_j, j \in J\}\right\}$,
$\mathbb{P}\left(\left(\frac{||\Pi_{S_{(k),(t)}} \mu||_n^2 + ||\Pi_{S_{(k),(t)}} \epsilon||_n^2}{\sigma^2} > x\right) \cap A_k\right)$

$$\begin{aligned}&= \mathbb{P}\left(\left(\frac{||\Pi_{S_{(k),(t)}} \epsilon||_n^2}{\sigma^2} > x\right) \cap A_k\right) \\ &\leq \mathbb{P}\left(\frac{||\Pi_{S_{(k),(t)}} \epsilon||_n^2}{\sigma^2} > x\right).\end{aligned}$$

And by construction of $U_{k,t}^1$,

$$\mathbb{P}\left(\frac{||\Pi_{S_{(k),(t)}} \epsilon||_n^2}{\sigma^2} > x\right) \leq \mathbb{P}\left(U_{k,t}^1 > x\right).$$

Thus

$$\mathbb{P}\left((U_{k,t} > x) \cap A_k\right) \leq \mathbb{P}\left(U_{k,t}^1 > x\right).$$

□

Proof of Theorem 3.2. Let $k < k_0$.
We use the identity $\forall (a,b) \in \mathbb{R}^2, (a+b)^2 \geq \frac{1}{2}a^2 - b^2$.
On the event A_{k_0}:
$\forall t \in I = \{0, ..., \log_2(k_0 - k)\}$:

$$\begin{aligned} ||\Pi_{S_{(k),(t)}} Y||_n^2 &= ||\Pi_{S_{(k),(t)}}(\mu + \epsilon)||_n^2 \\ &\geq \frac{1}{2}||\Pi_{S_{(k),(t)}}\mu||_n^2 - ||\Pi_{S_{(k),(t)}}\epsilon||_n^2 \\ &\geq \frac{1}{2}\inf\left\{||\Pi_S \mu||_n^2, S \in B_{2^t}\right\} - ||\Pi_{S_{(k),(t)}}\epsilon||_n^2 \end{aligned}$$

where $B_{2^t} = \{\text{span}(X_I), I \subset J, |I| = 2^t\}$. Hence:

$$\begin{aligned} &\mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}} Y||_n^2 \leq \overline{U_{k,t}^1}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right) \\ &= \mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}}(\mu+\epsilon)||_n^2 \leq \overline{U_{k,t}^1}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right) \\ &\leq \mathbb{P}\left(\forall t \in I, \frac{1}{2\sigma^2}\inf\left\{||\Pi_S\mu||_n^2, S \in B_{2^t}\right\} - \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}}\epsilon||_n^2 \leq \overline{U_{k,t}^1}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right). \end{aligned}$$

We have on the event A_{k_0} and for $k + 2^t \leq k_0$ that
$||\Pi_{S_{(k),(t)}}\epsilon||_n^2 \leq \sup\{||\Pi_S\epsilon||_n^2, S \in B_{2^t}\}$. Moreover, for $S \in B_{2^t}$, $||\Pi_S\epsilon||_n^2 \sim \frac{\sigma^2}{n}\chi_{2^t}^2$. Note that $|B_{2^t}| = \binom{k_0}{2^t}$.
Denote $Z_t = \frac{||\Pi_{S_{(k),(t)}}\epsilon||_n^2}{\sigma^2}$ and $\bar{Z}_t(u)$ the probability for the statistic Z_t to be larger than u. We denote $\bar{\chi}_d(u)$ the probability for a chi-square with d degrees of freedom to be larger than u. We have an upper bound of the $(1-u)$-quantile of the statistic Z_t: $\bar{Z}_t^{-1}(u) \leq \bar{\chi}_{2^t}^{-1}(u/|B_{2^t}|)/n$. Indeed:

$$\begin{aligned} \mathbb{P}\left(Z_t > \frac{\bar{\chi}_{2^t}^{-1}(u/|B_{2^t}|)}{n}\right) &\leq \mathbb{P}\left(\sup\left\{\frac{||\Pi_S\epsilon||_n^2}{\sigma^2}, S \in B_{2^t}\right\} > \frac{\bar{\chi}_{2^t}^{-1}(u/|B_{2^t}|)}{n}\right) \\ &\leq \sum_{S \in B_{2^t}} \mathbb{P}\left(||\Pi_S\epsilon||_n^2 > \frac{\sigma^2}{n}\bar{\chi}_{2^t}^{-1}(u/|B_{2^t}|)\right) \\ &\leq |B_{2^t}|\frac{u}{|B_{2^t}|} \leq u. \end{aligned}$$

Therefore, the following condition
$(cond_k) : \exists t \in I,$

$$\frac{1}{2\sigma^2}\inf\left\{||\Pi_S\mu||_n^2, S \in B_{2^t}\right\} \geq \frac{1}{n}\bar{\chi}_{2^t}^{-1}\left(\frac{\gamma/k_0}{|B_{2^t}|}\right) + \overline{U_{k,t}^1}^{-1}(\alpha_{k,t})$$

implies that:

$$\mathbb{P}\left[\forall t \in I, \frac{1}{2\sigma^2}\inf\left\{||\Pi_S\mu||_n^2, S \in B_{2^t}\right\} - \frac{||\Pi_{S_{(k),(t)}}\epsilon||_n^2}{\sigma^2} \leq \overline{U_{k,t}^1}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right] \leq \gamma/k_0. \quad (21)$$

Let us denote $\forall 0 < d$,

$$G_{k,d} = \{\text{span}(X_I), I \subset \{1,..,p\} \setminus \{(1),...,(k)\}, |I| = d\}. \tag{22}$$

Note that $|G_{k,d}| = \binom{p-k}{d}$.
Then $U^1_{k,t} \leq \sup\{||\Pi_S \epsilon||^2_n, S \in G_{k,2^t}\}$. This inequality leads us to an upper bound of the (1-u)-quantile of $U^1_{k,t}$:
$\overline{U^1_{k,t}}^{-1}(u) \leq \bar{\chi}^{-1}_{2^t}(u/|G_{k,2^t}|)/n$.

Using $\overline{U^1_{k,t}}^{-1}(u) \leq \bar{\chi}^{-1}_{2^t}(u/|G_{k,2^t}|)/n$ in the condition ($cond_k$), we obtain the following condition which still implies (21):

$$\exists t \in I, \frac{1}{2\sigma^2}\inf\left\{||\Pi_S \mu||^2_n, S \in B_{2^t}\right\} \geq \frac{1}{n}\left[\bar{\chi}^{-1}_{2^t}\left(\frac{\gamma/k_0}{|B_{2^t}|}\right) + \bar{\chi}^{-1}_{2^t}\left(\frac{\alpha_{k,t}}{|G_{k,2^t}|}\right)\right].$$

Moreover, Laurent and Massart (2000) showed that for $K \sim \chi^2_d$:

$$\forall x > 0, \mathbb{P}\left(K \geq d + 2\sqrt{dx} + 2x\right) \leq e^{-x}. \tag{23}$$

Then for $d = 2^t$ and $x_u = \log\left(\frac{|B_{2^t}|}{\gamma/k_0}\right)$ we have $\bar{\chi}^{-1}_{2^t}\left(\frac{\gamma/k_0}{|B_{2^t}|}\right) \leq 2^t + 2\sqrt{2^t x_u} + 2x_u$. Since $\binom{D}{d} \leq \left(\frac{eD}{d}\right)^d$, $|B_{2^t}| \leq \left(\frac{ek_0}{2^t}\right)^{2^t}$, thus $x_u = 2^t \log\left(\frac{ek_0}{2^t}\right) + \log\left(\frac{k_0}{\gamma}\right)$.
Using $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ for all $u > 0, v > 0$ and $\sqrt{u} \leq u$ for all $u \geq 1$, we obtain:

$$\begin{aligned}
\bar{\chi}^{-1}_{2^t}\left(\frac{\gamma/k_0}{|B_{2^t}|}\right) &\leq 2^t\left[1 + 2\sqrt{\log\left(\frac{ek_0}{2^t}\right)} + 2\log\left(\frac{ek_0}{2^t}\right)\right] \\
&\quad +2\left[\sqrt{2^t \log(k_0/\gamma)} + \log(k_0/\gamma)\right] \\
&\leq 2^t\left[5 + 4\log\left(\frac{k_0}{2^t}\right)\right] + 2\left[\sqrt{2^t\log(k_0/\gamma)} + \log(k_0/\gamma)\right].
\end{aligned}$$

For $d = 2^t$ and $x_u = \log(|G_{k,2^t}|/\alpha_{k,t})$, we obtain:
$\bar{\chi}^{-1}_{2^t}\left(\alpha_{k,t}/|G_{k,2^t}|\right)$

$$\begin{aligned}
&\leq 2^t\left[1 + 2\sqrt{\log\left(\frac{e(p-k)}{2^t}\right)} + 2\log\left(\frac{e(p-k)}{2^t}\right)\right] \\
&\quad +2\left[\sqrt{2^t\log(1/\alpha_{k,t})} + \log(1/\alpha_{k,t})\right] \\
&\leq 2^t\left[5 + 4\log\left(\frac{p-k}{2^t}\right)\right] + 2\left[\sqrt{2^t\log(1/\alpha_{k,t})} + \log(1/\alpha_{k,t})\right].
\end{aligned}$$

We also have an upper bound of $1/\alpha_{k,t}, \forall t \in \mathcal{T}_k$. Indeed , the construction of $\{\alpha_{k,t}, t \in \mathcal{T}_k\}$ with the procedure P3 gives $\mathbb{P}\left(\exists t \in \mathcal{T}_k, U^1_{k,t} > \overline{U^1_{k,t}}^{-1}(\alpha_{k,t})\right) = \alpha$. Thus $\forall t \in \mathcal{T}_k, \alpha_{k,t} \geq$

$\alpha/|\mathcal{T}_k|$, since $\mathbb{P}\left(\exists t \in \mathcal{T}_k, U^1_{k,t} > \overline{U^1_{k,t}}^{-1}(\alpha/|\mathcal{T}_k|)\right) \leq \alpha$.
Hence we obtain:

$$\begin{aligned}\bar{\chi}^{-1}_{2^t}\left(\alpha_{k,t}/|G_{k,2^t}|\right) &\leq 2^t\left[5+4\log\left(\frac{p-k}{2^t}\right)\right] \\ &\quad +2\left[\sqrt{2^t\log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)}+\log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)\right].\end{aligned}$$

Using the inequality $a\sqrt{u}+b\sqrt{v} \leq \sqrt{a^2+b^2}\sqrt{u+v}$ which holds for any positive numbers a, b, u, v, we finally get the condition $(R_{2,k})$ which implies (21):
$(R_{2,k})$: $\exists t \in I$ such that

$$\begin{aligned}\frac{1}{2\sigma^2}\inf\left\{||\Pi_S\mu||^2_n, S \in B_{2^t}\right\} &\geq \frac{2^t}{n}\left[10+4\log\left(\frac{(p-k)k_0}{2^{2t}}\right)\right] \\ &+ \frac{2}{n}\left[\sqrt{2^{t+1}\log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)}+\log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)\right].\end{aligned}$$

This leads to

$$\mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}}Y||^2_n \leq \overline{U^1_{k,t}}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right) \leq \gamma/k_0.$$

Hence

$$\mathbb{P}\left(\forall t \in I, U_{k,t} \leq \overline{U^1_{k,t}}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right) \leq \gamma/k_0.$$

Then, $\forall k < k_0, \mathbb{P}\left(\mathring{k}_A = k \ \cap \ A_{k_0}\right) \leq \gamma/k_0$.

We can calculate $\mathbb{P}_\mu(\hat{J} \neq J)$:

$$\begin{aligned}\mathbb{P}_\mu(\hat{J} \neq J) &\leq \mathbb{P}_\mu(\hat{J} \neq J \ \cap A_{k_0}) + \mathbb{P}(A^c_{k_0}) \\ &\leq \left(\sum_{j=0}^{k_0-1}\mathbb{P}_\mu(\mathring{k}_A = j \ \cap A_{k_0}) + \mathbb{P}_\mu(\mathring{k}_A > k_0 \ \cap A_{k_0})\right) + \mathbb{P}(A^c_{k_0}) \\ &\leq k_0\gamma/k_0 + \alpha + \delta.\end{aligned}$$

And then (13) is proved. □

Proof of Lemma A.1. Under $\hat{H}_k$ and on the event A_k :

$$\begin{aligned}U_{k,t} &= ||\Pi_{S_{(k),(t)}}Y||^2_n/\sigma^2 = ||\Pi_{S_{(k),(t)}}(\mu+\epsilon)||^2_n/\sigma^2 \\ &= ||\Pi_{S_{(k),(t)}}\epsilon||^2_n/\sigma^2.\end{aligned}$$

The family $(X_i)_i$ is orthonormal, thus:
$U_{k,t} = \sum_{j=k+1}^{k+2^t} < \epsilon, X_{(j)} >^2_n /\sigma^2$.
As $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$, we have for all $1 \leq j \leq p$,
$< \epsilon, X_j > \sim \mathcal{N}(0, \sigma^2)$ and the variables $< \epsilon, X_j >, j = 1, ..., p$ are i.i.d.. Thus $\left\{< \epsilon, X_{(j)} >, j > k\right\} = \{< \epsilon, X_m >, m \notin J\} = \{\sigma W_1, ..., \sigma W_{p-k}\}$.

So $\sum_{j=k+1}^{k+2^t} < \epsilon, X_{(j)} >^2_n /\sigma^2 \leq \sum_{j=1}^{2^t} W^2_{(j)}/n = Z_{k,D_{k,t}}/n$.

□

Proof of Corollary A.2. Let $k < k_0$.
σ_2 is defined such that $|\beta_{\sigma_2(1)}| \leq ... \leq |\beta_{\sigma_2(k_0)}|$, note $\epsilon_{(j+1)} = ||\Pi_{S_{(j),0}}\epsilon||$, $\forall j \in \{k+1, ..., k+2^t\}$ with $k+2^t \leq k_0$.
Similarly as in the proof of Theorem 3.2, using that
$\inf\{||\Pi_S\mu||_n^2, S \in B_{2^t}\} = \sum_{j=1}^{2^t} \beta_{\sigma_2(j)}^2$, we get that:

$$\mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}}Y||_n^2 \leq \frac{\bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_{k,t})}{n} \cap A_{k_0}\right)$$

$$\leq \mathbb{P}\left(\forall t \in I, \frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta_{\sigma_2(j)}^2 - \frac{1}{n\sigma^2}\sum_{j=k}^{k+2^t-1}\epsilon_{(j+1)}^2 \leq \frac{\bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_{k,t})}{n} \cap A_{k_0}\right).$$

On the event A_{k_0}, $\left\{< \epsilon, X_{(j+1)} >, k \leq j \leq k+2^t-1\right\} \subset \{< \epsilon, X_j >, j \in J\}$, which implies that we have a stochastic upper bound: $\sum_{j=k}^{k+2^t-1}\epsilon_{(j+1)}^2 \leq \sigma^2 Z_{2^t,k_0}$.
Hence the following condition

$$\exists\, t \leq \log_2(k_0-k)/\frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta_{\sigma_2(j)}^2 \geq \frac{1}{n}\left[\bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_{k,t}) + \bar{Z}^{-1}_{D_{k,t},k_0}(\gamma/k_0)\right] \tag{24}$$

implies that

$$\mathbb{P}\left(\forall t \in I, \frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta_{\sigma_2(j)}^2 - \frac{1}{n\sigma^2}\sum_{j=k}^{k+2^t-1}\epsilon_{(j+1)}^2 \leq \frac{\bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_{k,t})}{n} \cap A_{k_0}\right)$$

$$\leq \gamma/k_0.$$

This leads to

$$\mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}}Y||_n^2 \leq \bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_{k,t}) \cap A_{k_0}\right) \leq \gamma/k_0. \tag{25}$$

Let $0 < u < 1$, $0 < D$ and $d < D$. In the following, we study the behavior of the $(1-u)$ quantile of the statistic $Z_{d,D}$ in order to obtain a more explicit condition than (24).
Let define $V_{d,D} = \{I \subset \{1, ..., D\}/|I| = d\}$. Note that $|V_{d,D}| = \binom{D}{d}$. Let recall that $Z_{d,D}$ is defined by (18) as $Z_{d,D} = \sum_{j=1}^{d} W_{(j)}^2$ where $W_1, ..., W_D$ are D i.i.d. standard Gaussian variables ordered as $|W_{(1)}| \geq ... \geq |W_{(D)}|$.
We have that: $Z_{d,D} \leq \sup\left\{\sum_{i\in I} W_i^2, I \in V_{d,D}\right\}$. Note that for $I \in V_{d,D}$, $\sum_{i\in I} W_i^2, \sim \chi_d^2$.
We obtain that the $(1-u)$-quantile of $Z_{d,D}$ is lower than $\bar{\chi_d}^{-1}(u/|V_{d,D}|)$:

$$\begin{aligned}
\mathbb{P}\left(Z_{d,D} > \bar{\chi_d}^{-1}(u/|V_{d,D}|)\right) &\leq \mathbb{P}\left(\sup_{I\in V_{d,D}}(\sum_{i\in I} W_i^2) > \bar{\chi_d}^{-1}(u/|V_{d,D}|)\right) \\
&\leq \sum_{I\in V_{d,D}} \mathbb{P}\left(\sum_{i\in I} W_i^2 > \bar{\chi_d}^{-1}(u/|V_{d,D}|)\right) \\
&\leq |V_{d,D}|\frac{u}{|V_{d,D}|} \leq u.
\end{aligned}$$

Using the expression of the upper bound of $\bar{\chi}_d^{-1}(u)$ from the proof of Theorem 3.2, we get the condition $(R_{2bis,k})$ from an upper bound of the right part in the condition (24). The end of the proof is the same as in the proof of Theorem 3.2. □

Proof of Lemma 3.3. Let $x > 0$, $1 \leq k < \min(n-1, p)$ and $t \in \mathcal{T}_k$. By definition of $\tilde{U}_{D_{k,t},N_{k,t}}$, we have

$$\begin{aligned}\mathbb{P}\left((\tilde{U}_{D_{k,t},N_{k,t}} > x) \cap A_k\right) &= \mathbb{P}\left(\left(\frac{N_{k,t}||\Pi_{S_{(k),(t)}}Y||_n^2}{D_{k,t}||Y - \Pi_{V_{(k),(t)}}Y||_n^2} > x\right) \cap A_k\right) \\ &= \mathbb{P}\left(\left(\frac{N_{k,t}||\Pi_{S_{(k),(t)}}\mu||_n^2 + N_{k,t}||\Pi_{S_{(k),(t)}}\epsilon||_n^2}{D_{k,t}||\mu + \epsilon - \Pi_{V_{(k),(t)}}\mu - \Pi_{V_{(k),(t)}}\epsilon||_n^2} > x\right) \cap A_k\right).\end{aligned}$$

Since $A_k = \left\{\left\{X_{(1)}, \ldots, X_{(k)}\right\} = \{X_j, j \in J\}\right\}$,

$$\begin{aligned}&\mathbb{P}\left(\left(\frac{N_{k,t}||\Pi_{S_{(k),(t)}}\mu||_n^2 + N_{k,t}||\Pi_{S_{(k),(t)}}\epsilon||_n^2}{D_{k,t}||\mu + \epsilon - \Pi_{V_{(k),(t)}}\mu - \Pi_{V_{(k),(t)}}\epsilon||_n^2} > x\right) \cap A_k\right) \\ &\qquad = \mathbb{P}\left(\left(\frac{N_{k,t}||\Pi_{S_{(k),(t)}}\epsilon||_n^2}{D_{k,t}||\epsilon - \Pi_{V_{(k),(t)}}\epsilon||_n^2} > x\right) \cap A_k\right) \\ &\qquad \leq \mathbb{P}\left(\frac{N_{k,t}||\Pi_{S_{(k),(t)}}\epsilon||_n^2}{D_{k,t}||\epsilon - \Pi_{V_{(k),(t)}}\epsilon||_n^2} > x\right).\end{aligned}$$

And by construction of $\Upsilon_{k,t}$,

$$\mathbb{P}\left(\frac{N_{k,t}||\Pi_{S_{(k),(t)}}\epsilon||_n^2}{D_{k,t}||\epsilon - \Pi_{V_{(k),(t)}}\epsilon||_n^2} > x\right) \leq \mathbb{P}\left(\Upsilon_{k,t} > x\right).$$

Thus

$$\mathbb{P}\left((\tilde{U}_{D_{k,t},N_{k,t}} > x) \cap A_k\right) \leq \mathbb{P}\left(\Upsilon_{k,t} > x\right).$$

□

Proof of Theorem 3.4. Let $k < k_0$ and $0 < \gamma < 1$. Denote $I = \{0, \ldots, \lfloor \log_2(k_0 - k) \rfloor\}$. From the proof of Theorem 3.2 (more precisely the condition $(cond_k)$), we have that if the following condition is verified:
$\exists t \in I$ such that

$$\frac{1}{2}\inf\left\{||\Pi_S\mu||_n^2, S \in B_{2^t}\right\} \geq \bar{\Upsilon}_{k,t}^{-1}(\alpha_{k,t})Q_{1-\gamma/2k_0}\frac{D_{k,t}}{N_{k,t}} + \frac{\sigma^2}{n}\bar{\chi}_{2^t}^{-1}\left(\frac{\gamma/2k_0}{|B_{2^t}|}\right), \tag{26}$$

where Q_{1-u} denote the $(1-u)$-quantile of the statistics $||Y - \Pi_{V_{(k),(t)}}Y||_n^2$ under the event A_{k_0},
then we have:

$$\mathbb{P}\left(\forall t \in I, ||\Pi_{S_{(k),(t)}}Y||_n^2 \leq \bar{\Upsilon}_{k,t}^{-1}(\alpha_{k,t})Q_{1-\gamma/2k_0}\frac{D_{k,t}}{N_{k,t}} \cap A_{k_0}\right) \leq \gamma/2k_0.$$

Since
$\mathbb{P}\left(\forall t \in I, \tilde{U}_{D_{k,t},N_{k,t}} < \bar{\Upsilon}_{k,t}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right)$

$$\leq \inf_{t\in I}\left\{\mathbb{P}\left(\tilde{U}_{D_{k,t},N_{k,t}} < \bar{\Upsilon}_{k,t}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right)\right\}$$

and since
$\mathbb{P}\left(\tilde{U}_{D_{k,t},N_{k,t}} < \bar{\Upsilon}_{k,t}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right)$

$$\begin{aligned}
&\leq \underbrace{\mathbb{P}\left(||Y-\Pi_{V_{(k),(t)}}Y||_n^2 > Q_{1-\gamma/2k_0} \ \cap A_{k_0}\right)}_{\leq\gamma/2k_0} \\
&\quad +\mathbb{P}\left(\frac{||\Pi_{S_{(k),(t)}}Y||_n^2}{D_{k,t}} \leq \bar{\Upsilon}_{k,t}^{-1}(\alpha_{k,t})\frac{Q_{1-\gamma/2k_0}}{N_{k,t}} \ \cap A_{k_0}\right) \\
&\leq \gamma/k_0,
\end{aligned}$$

we have that the condition (26) implies that

$$\mathbb{P}\left(\forall t \in I, \tilde{U}_{D_{k,t},N_{k,t}} < \bar{\Upsilon}_{k,t}^{-1}(\alpha_{k,t}) \ \cap A_{k_0}\right) \leq \gamma/k_0. \tag{27}$$

In the following, we give an upper bound of the right part in (26). For this doing, we have to give an upper bound of $\bar{\Upsilon}_{k,t}^{-1}(\alpha_{k,t})$ and $Q_{1-\gamma/2k_0}$.

Assume we are on the event A_k, then

$$\begin{aligned}
\Upsilon_{k,t} &= \frac{N_{k,t}||\Pi_{S_{(k),\sigma_1(t)}}Y||_n^2}{D_{k,t}||Y-\Pi_{V_{(k),\sigma_1(t)}}Y||_n^2} \\
&= \frac{N_{k,t}||\Pi_{S_{(k),\sigma_1(t)}}\epsilon||_n^2}{D_{k,t}||Y-\Pi_{V_{(k)}}Y-\Pi_{S_{(k),\sigma_1(t)}}\epsilon||_n^2}.
\end{aligned}$$

As we are on the event A_k, the space $V_{(k)}$ is not a random space. Thus for any subspaces S of dimension $D_{k,t} = 2^t$, we have that $||\Pi_S Y||_n^2 = ||\Pi_S \epsilon||_n^2 \sim \sigma^2\chi^2_{2^t}/n$ and we have that $||Y-\Pi_{V_{(k)}}Y-\Pi_S Y||_n^2 = ||\Pi_{(S\oplus V_{(k)})^\perp}\epsilon||_n^2 \sim \sigma^2\chi^2_{n-(2^t+k)}/n$.
Hence $\dfrac{N_{k,t}||\Pi_S Y||_n^2}{D_{k,t}||Y-\Pi_{V_{(k)}}Y-\Pi_S Y||_n^2} \sim F_{D_{k,t},N_{k,t}}$. Thus on the
event A_k, $\Upsilon_{k,t} \leq \sup\left\{\dfrac{N_{k,t}||\Pi_S\epsilon||_n^2}{D_{k,t}||\epsilon-\Pi_{V_{(k)}+S}\epsilon||_n^2}, S \in G_{k,2^t}\right\}$,
where $G_{k,2^t}$ is defined by (22).
We deduce that the $(1-u)$-quantile of $\Upsilon_{k,t}$ is lower that
$\bar{F}^{-1}_{D_{k,t},N_{k,t}}(u/|G_{k,2^t}|)$. Indeed:

$$\begin{aligned}
&\mathbb{P}\left(\Upsilon_{k,t} > \bar{F}^{-1}_{D_{k,t},N_{k,t}}\left(u/|G_{k,2^t}|\right)\right) \\
&\qquad \leq \mathbb{P}\left[\sup\left\{\frac{N_{k,t}||\Pi_S\epsilon||_n^2}{D_{k,t}||\epsilon-\Pi_{V_{(k)}+S}\epsilon||_n^2}, S\in G_{k,2^t}\right\}\right. \\
&\qquad\qquad \left. > \bar{F}^{-1}_{D_{k,t},N_{k,t}}\left(u/|G_{k,2^t}|\right)\right] \\
&\qquad \leq \sum_{S\in G_{k,2^t}} \mathbb{P}\left(\frac{N_{k,t}||\Pi_S\epsilon||_n^2}{D_{k,t}||\epsilon-\Pi_{V_{(k)}+S}\epsilon||_n^2} > \bar{F}^{-1}_{D_{k,t},N_{k,t}}\left(u/|G_{k,2^t}|\right)\right) \\
&\qquad \leq |G_{k,2^t}|\frac{u}{|G_{k,2^t}|} \leq u.
\end{aligned}$$

Baraud et al. (2003) gave an upper bound of $\bar{F}^{-1}_{D,N}(u)$, for $0<D$, $0<N$ and $0<u$:

$$\begin{aligned}
D\bar{F}^{-1}_{D,N}(u) &\leq D + 2\sqrt{D\left(1+\frac{D}{N}\right)\log\left(\frac{1}{u}\right)} \\
&\quad + \left(1+2\frac{D}{N}\right)\frac{N}{2}\left[\exp\left(\frac{4}{N}\log\left(\frac{1}{u}\right)\right)-1\right].
\end{aligned}$$

Since $\exp(u)-1 \leq u\exp(u)$ for any $u>0$, $\sqrt{u+v} \leq \sqrt{u}+\sqrt{v}$ for all $u>0, v>0$ and since $\alpha_{k,t} \geq \alpha/|\mathcal{T}_k|$, we derive that:

$$\begin{aligned}
2^t\bar{\Upsilon}^{-1}_{k,t}(\alpha_{k,t}) &\leq 2^t\left[1+\Lambda_3(k,t)\log\left(\frac{e(p-k)}{2^t}\right)\right] \\
&\quad + 2\left[\sqrt{2^t\left(1+\frac{2^t}{N_{k,t}}\right)\log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)} + \frac{\Lambda_2(k,t)}{2}\log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)\right],
\end{aligned}$$

where $\Lambda_1(k,t) = \sqrt{1+\frac{D_{k,t}}{N_{k,t}}}$, $\Lambda_2(k,t) = \left(1+2\frac{D_{k,t}}{N_{k,t}}\right)M$ and $\Lambda_3(k,t) = 2\Lambda_1(k,t)+\Lambda_2(k,t)$ with $L_t = \log(|\mathcal{T}_k|/\alpha)$, $m_t = \exp(4L_t/N_{k,t})$, $m_p = \exp\left(\frac{4D_{k,t}}{N_{k,t}}\log\left(\frac{e(p-k)}{2^t}\right)\right)$, $M = 2m_tm_p$.
Since $\sqrt{ab}+mb \leq a/2+(m+1/2)b$ holds for any positive numbers a, b, m, we obtain that:

$$\begin{aligned}
2^t\bar{\Upsilon}^{-1}_{k,t}(\alpha_{k,t}) &\leq 2^t\left[1+\Lambda_1^2(k,t)+\Lambda_3(k,t)\log\left(\frac{e(p-k)}{2^t}\right)\right] &(28)\\
&\quad + (1+\Lambda_2(k,t))\log\left(\frac{|\mathcal{T}_k|}{\alpha}\right). &(29)
\end{aligned}$$

We have now to find an upper bound of $Q_{1-\gamma/2k_0}$.
$Q_{1-\gamma/2k_0}$ is defined by $\mathbb{P}\left(||Y-\Pi_{V_{(k),(t)}}Y||_n^2 > Q_{1-\gamma/2k_0} \cap A_{k_0}\right) \leq \gamma/2k_0$.
We always have that: $||Y-\Pi_{V_{(k),(t)}}Y||_n^2 \leq ||\mu||_n^2+||\epsilon||_n^2$. Thus $\forall\ 0<u<1$, the $(1-u)$-quantile of $||Y-\Pi_{V_{(k),(t)}}Y||_n^2$ is lower than the $(1-u)$-quantile of $||\mu||_n^2+||\epsilon||_n^2$.
As $||\epsilon||_n^2 \sim \sigma^2\chi_n^2/n$, we can use the equation (23) for $x_u = \log(2k_0/\gamma)$ and we obtain that

$\bar{\chi}_n^{-1}(\gamma/2k_0) \leq n + 2\sqrt{nx_u} + 2x_u$.
Therefore

$$Q_{1-\gamma/2k_0} \leq ||\mu||_n^2 + \sigma^2 \frac{n + 2\sqrt{nx_u} + 2x_u}{n} \tag{30}$$

and as $1 + 2\sqrt{u} + 2u \leq 2 + 3u$, we get

$$Q_{1-\gamma/2k_0} \leq ||\mu||_n^2 + \sigma^2 \left(2 + \frac{3}{n} \log\left(\frac{2k_0}{\gamma}\right)\right). \tag{31}$$

Combining (28), (31) in (26) and using that

$$\begin{aligned} \bar{\chi}_{2^t}^{-1}\left(\frac{\gamma/2k_0}{|B_{2^t}|}\right) &\leq 2^t\left[5 + 4\log\left(\frac{k_0}{2^t}\right)\right] + 2\left[\sqrt{2^t \log(2k_0/\gamma)} + \log(2k_0/\gamma)\right] \\ &\leq 2^t\left[6 + 4\log\left(\frac{k_0}{2^t}\right)\right] + 3\log(2k_0/\gamma), \end{aligned}$$

we obtain the following condition:
$(R_{3,k})$: $\exists t \in I$ such that
$\frac{1}{2}\inf\{||\Pi_S \mu||_n^2, S \in B_{2^t}\}$

$$\begin{aligned} &\geq \frac{D_{k,t}\bar{F}^{-1}_{D_{k,t},N_{k,t}}(\alpha_{k,t}/|G_{2^t}|)}{N_{k,t}}\left[||\mu||_n^2 + \sigma^2\left(2 + \frac{3}{n}\log\left(\frac{2k_0}{\gamma}\right)\right)\right] \\ &\quad + \frac{\sigma^2}{n}\left[2^t\left(6 + 4\log\left(\frac{k_0}{2^t}\right)\right) + 3\log\left(\frac{2k_0}{\gamma}\right)\right] \\ &\geq \frac{A(k,t)}{N_{k,t}}\left[||\mu||_n^2 + \sigma^2\left(2 + \frac{3}{n}\log\left(\frac{2k_0}{\gamma}\right)\right)\right] \\ &\quad + \frac{\sigma^2}{n}\left[2^t\left(6 + 4\log\left(\frac{k_0}{2^t}\right)\right) + 3\log\left(\frac{2k_0}{\gamma}\right)\right], \end{aligned}$$

where $A(k,t) = 2^t\left[2 + \frac{2^t}{N_{k,t}} + \Lambda_3(k,t)\log\left(\frac{e(p-k)}{2^t}\right)\right] + (1 + \Lambda_2(k,t))\log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)$.

The condition $(R_{3,k})$ leads to (27) and thus

$$\begin{aligned} \mathbb{P}_\mu(\hat{J} \neq J) &\leq \mathbb{P}_\mu(\hat{J} \neq J \ \cap A_{k_0}) + \mathbb{P}(A_{k_0}^c) \\ &\leq \left(\sum_{j=0}^{k_0-1} \mathbb{P}_\mu(\mathring{k}_B = j \ \cap A_{k_0}) + \mathbb{P}_\mu(\mathring{k}_B > k_0 \ \cap A_{k_0})\right) + \mathbb{P}(A_{k_0}^c) \\ &\leq k_0\gamma/k_0 + \alpha + \delta. \end{aligned}$$

And then (15) is proved. □

Proof of Remark 3.5. In the following, $C(a,b)$ denote a constant depending on the parameters a and b. Under the assumption that $2^t \leq (n-k)/2$ and since $\forall\, x \geq 2$, $\frac{\log(x)}{x} \leq 1$ we have that:

$$\frac{D_{k,t}}{N_{k,t}}\log\left(\frac{p-k}{D_{k,t}}\right) \leq \frac{2^t}{n-k-2^t}\log\left(\frac{n-k}{2^t}\right) \leq 2\frac{2^t}{n-k}\log\left(\frac{n-k}{2^t}\right) \leq 2.$$

Moreover the ratio $D_{k,t}/N_{k,t}$ is bounded by 1, thus $\log(m_p) \leq 4\frac{D_{k,t}}{N_{k,t}} + 4\frac{D_{k,t}}{N_{k,t}} \log\left(\frac{p-k}{D_{k,t}}\right) \leq 12$.
As the ratio $4L_{k,t}/N_{k,t}$ is bounded by $C'(\alpha)$ and since
$M \leq 2\exp(C'(\alpha))\exp(12)$, we have that M is bounded by $C''(\alpha)$. Thus $\Lambda_1(k,t) \leq \sqrt{2}$, $\Lambda_2(k,t) \leq 3C''(\alpha)$ and $\Lambda_3(k,t) \leq 2\sqrt{2} + 3C''(\alpha)$.
We obtain under the condition $\log(p-k) > 1$ that $A(k,t) \leq 2^t C(\alpha)\log(p-k)$.
We also have that $\left[||\mu||_n^2 + \sigma^2\left(2 + \frac{3}{n}\log\left(\frac{2k_0}{\gamma}\right)\right)\right] \leq C(||\mu||_n, \gamma, \sigma)$
since $\log(k_0)/n \leq 1$, and that
$2^t\left(6 + 4\log\left(\frac{k_0}{2^t}\right)\right) + 3\log\left(\frac{2k_0}{\gamma}\right) \leq 2^t\left[6 + 4\log\left(\frac{k_0}{2^t}\right) + 3\log\left(\frac{2k_0}{\gamma}\right)\right] \leq 2^t C(\gamma)\log(k_0)$.

We finally obtain equation (16). □

3.3 Simulations et données réelles

3.3.1 Résultats de simulations

Reprenons l'exemple de la Table 3 et appliquons notre procédure de sélection ordonnée ainsi que notre procédure de sélection non ordonnée à variance inconnue. Les résultats sont présentés dans la Table 4, où "proc-ordered" fait référence à la procédure de sélection ordonnée, "procpval" , "proclas" et "procbol" à la méthode de sélection non ordonnée à variance inconnue lorsque la première étape -ordre des variables- est réalisée avec un ordre par p-valeurs, un ordre avec le chemin de régularisation du Lasso ou un ordre avec le chemin de régularisation du Bolasso, respectivement.

	Idéal	proc-ordered		procpval		proclas		procbol	
		α=0.1	α=0.05	α=0.1	α=0.05	α=0.1	α=0.05	α=0.1	α=0.05
				$\hat{\delta} = 1.00$		$\hat{\delta} = 1.00$		$\hat{\delta} = 0.17$	
Egalité	1.00	0.89	0.95	0.00	0.00	0.00	0.00	0.83	0.83
Incl.	11.00	11.66	11.21	5.88	5.36	9.82	9.32	11.30	11.20
C. incl.	11.00	11.00	11.00	5.68	5.23	8.15	7.83	10.99	10.99
MSE	0.00	0.12	0.11	4.11	4.56	2.12	2.40	0.11	0.11

TABLE 4 – **Résultats de 500 simulations pour un modèle dans lequel $n = 100, p = 600, k_0 = 11, \beta_J = 10$. δ est une estimation de la probabilité de se tromper en ordonnant les variables. La deuxième ligne "Egalité" donne le pourcentage de fois où $\hat{J} = J$. "Incl." donne la moyenne du nombre de variables sélectionnées et "C. incl." celle du nombre de variables pertinentes sélectionnées. Le MSE est obtenu par moyenne sur toutes les simulations : $MSE = \sum_{i=1}^{n}(\hat{Y}_i - (X\beta_J)_i)^2/n$, où $\hat{Y} = X\hat{\beta}$, et où $\hat{\beta}$ est une estimation de β avec des coefficients non nuls seulement sur $\hat{J}$.**

La procédure ordonnée donne de très bons résultats ainsi que la méthode "procbol". Comme mentionné dans la section précédente, l'ordonnancement des variables est très important, c'est là que se situe la différence entre "procpval" "proclas" et "procbol", différence qui se ressent énormément dans les résultats. En effet, d'après la Table 4 ordonner les variables à l'aide des p-valeurs, ou du chemin de régularisation du Lasso, a une probabilité de 1.00 de donner un mauvais résultat sur ce modèle simulé, c'est-à-dire de considérer au moins une variable non pertinente plus importante qu'une variable pertinente ; ce qui signifie que quelque soit la méthode de sélection de variables basée sur cet ordre, elle ne donnera pas de bons résultats en terme d'estimation exacte du support de β. En ce qui concerne la méthode d'ordonnancement des variables obtenues à partir du chemin de régularisation du Bolasso, on observe une probabilité de mal ordonner les variables de 0.17, ce qui est très raisonnable pour un cas à la limite de la très grande dimension ($\frac{k}{n}ln(\frac{p}{k}) = 0.44$) ; cet ordre permet ainsi d'avoir de très bon résultats dans l'estimation de J par notre méthode

de tests multiples.
La Figure 6 nous montre, sur le modèle simulé dans la Table 4, les variations de la probabilité (estimée sur 100 simulations) de mal ordonner les variables en fonction du nombre de bootstrap effectué . On confirme que pour la méthode Lasso (qui correspond à $m = 1$) on obtient $\delta = 1.00$. On observe que la probabilité de mal ordonner les variables décroît rapidement lorsque le nombre d'échantillons bootstrap augmente.

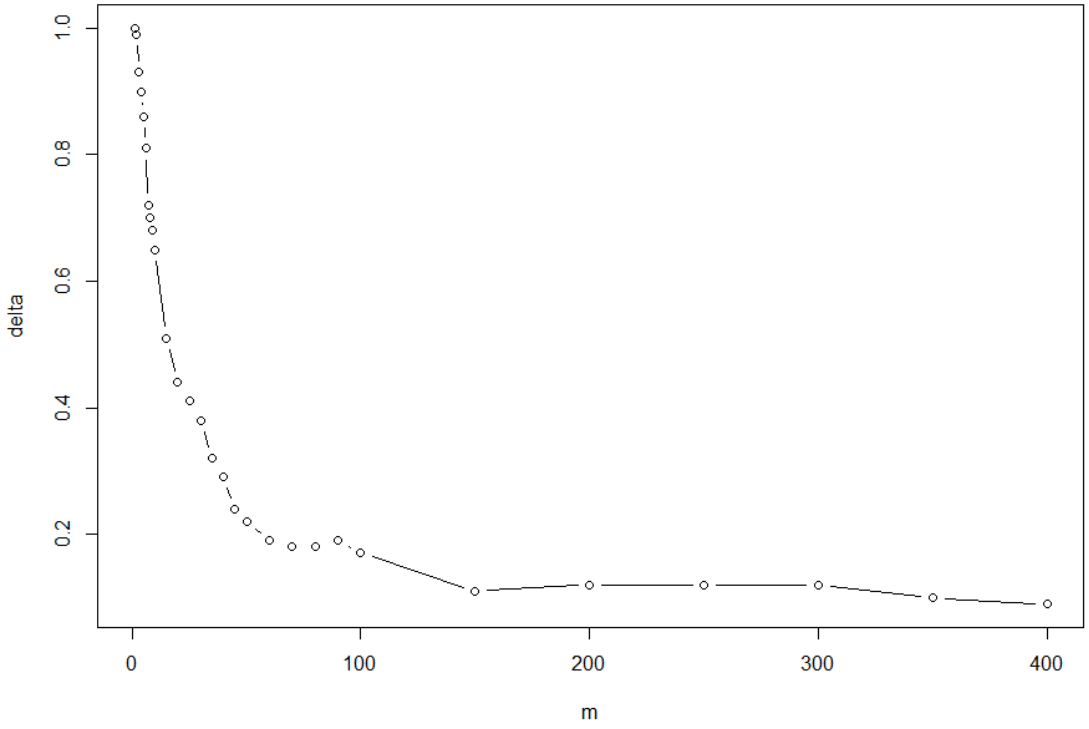


FIGURE 6 – **Probabilité de mal ordonner les variables à l'aide de la technique Bolasso en fonction du nombre m d'échantillons bootstrap. Le modèle utilisé est $n = 100, p = 600, k_0 = 11, \beta_J = 10$.**

Cette simulation ainsi que celles présentées dans la Section 3.2 ont été réalisées à l'aide du package R 'mht' pour 'multiple hypotheses testing for variable selection'. Ce package contient nos procédures de sélection de variables ainsi que la méthode Bolasso, il est disponible sur le CRAN (http ://cran.R-project.org).

3.3.2 Application aux données réelles

La procédure de tests multiples à variance inconnue pour la sélection non ordonnée "procbol" a été appliquée aux données réelles afin d'expliciter les relations biologiques potentielles entre les données métabolomiques et les données phénotypiques. Afin de faciliter l'interprétation biologique des variables sélectionnées, la procédure est appliquée sur les données brutes. On se focalise sur les phénotypes "LMP" et "DFI" qui font partie de la liste des phénotypes où l'apport des données métabolomiques est non négligeable d'après la Section 2.4 et la Figure 4. Les variables les plus souvent sélectionnées sur 100 itérations bootstrap pour les 3 modèles étudiés dans la Partie 2 sont reportées dans la Table 5.

Les métabolites sélectionnés par notre procédure de tests multiples dans le modèle 1 (modèle qui contient uniquement les données métabolites, cf. (12a)) pour le phénotype LMP sont aussi sélectionnés par la méthode Lasso sur ce même modèle cf. Section 2.2, mais en moindre quantité, ce qui tendrait à confirmer le comportement de ces deux méthodes observé sur les simulations : le Lasso sélectionne beaucoup de variables (dont un certain nombre se révèle être non pertinent en simulations) et la méthode de tests beaucoup moins. On peut donc considérer que les variables fortement sélectionnées par le Lasso mais qui ne le sont pas du tout par la méthode de tests multiples sont en réalité des variables non pertinentes. Cependant, cette conclusion est à mettre en balance avec de meilleures erreurs de prédiction pour la méthode Lasso. En effet, la procédure de tests multiples offre des résultats plus stables, cf. Figure 7(a) pour le phénotype DFI où l'on peut voir que le Lasso sélectionne beaucoup de variables avec des occurrences à plus de 50 contrairement à notre procédure de tests, mais elle donne des résultats de prédiction moins bons que ceux du Lasso, cf. Figure 7(b). Néanmoins la procédure de tests est destinée à sélectionner les variables pertinentes et non à faire de la prédiction. Il est à noter que, contrairement aux résultats sur les données réelles, le MSE est toujours meilleur sur nos simulations pour la méthode procbol que pour la méthode Lasso, Table 3 vs Table 4. Une raison possible aux résultats observés sur les données réelles est que l'hypothèse de parcimonie n'est pas vérifiée sur ces données. Par ailleurs l'hypothèse d'indépendance entre les observations est également contestable de part les relations de parenté entre les individus. Nous allons donc introduire dans le chapitre suivant la sélection de variables dans les modèles linéaires mixtes.

DFI					
Model 1		Model 2		Model 3	
δ(ppm) (n)	Assignement	δ(ppm) (n)	Assignement	δ(ppm) (n)	Assignement
4.05 (100)	creatinine	4.05 (96)	creatinine	4.05 (100)	creatinine
2.43 (86)	glutamine				
1.47 (81)	?				
2.51 (64)	citrate				
3.02 (63)	?				
LMP					
Model 1		Model 2		Model 3	
δ(ppm) (n)	Assignement	δ(ppm) (n)	Assignement	δ(ppm) (n)	Assignement
4.05 (100)	creatinine	4.05 (100)	creatinine	4.05 (100)	creatinine
3.93 (100)	creatine				
2.43 (93)	glutamine				
2.25 (61)	valine				

TABLE 5 – **Variables sélectionnées pour les phénotypes "DFI" et "LMP" à partir des données métabolomiques brutes sur les 3 modèles (12). Le décalage chimique (δ) en ppm est donné. Le nombre de fois où la variable est sélectionnée sur les 100 itérations est donné entre parenthèses, seuillé à 60.**

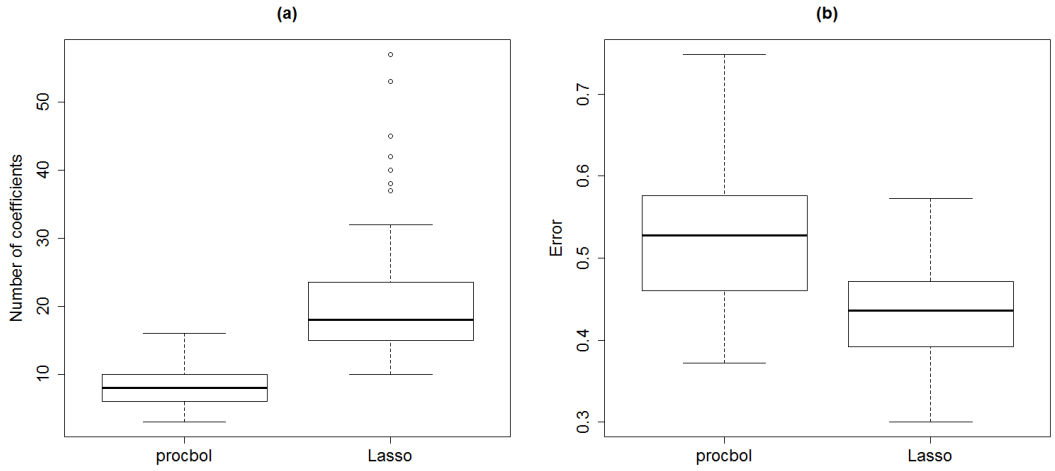


FIGURE 7 – **Nombre de coefficients sélectionnés et erreurs de prédiction pour le phénotype DFI et les méthodes Lasso et procbol sur le modèle ne contenant que les données métabolomiques (modèle 1, cf. (12a).)**

4 Sélection des effets fixes et aléatoires dans un modèle linéaire mixte

4.1 Motivations

Toutes les méthodes présentées ont été appliquées au jeu de données réelles présenté dans la Section 1.2. Les résultats étant peu concluants pour certains phénotypes étudiés, le travail de recherche s'est ensuite porté sur une méthode différente permettant d'intégrer toute l'information disponible sur le jeu de données. En effet, les animaux possèdent des liens de parenté qui peuvent influer sur les résultats, certains étant demi-frères. Les animaux ont aussi été élevés par lots, certains ont donc été soumis aux mêmes conditions environnementales, et ces conditions sont connues pour influencer les données métabolomiques ainsi que certains phénotypes. Dans les modèles précédents nous avons pris en compte ces variables au même titre que les variables métabolomiques. Cependant, si l'on considère que ces variables sont en fait des variables gaussiennes ayant une variance inconnue, alors on se place dans le cadre du modèle linéaire mixte, cf. Section 1.4, qui est tout adapté aux observations répétées. Les problèmes présents dans le modèle linéaire tels que la grande dimension ou le sur-apprentissage sont aussi présents dans un modèle linéaire mixte. L'objectif reste donc le même que dans la section précédente : identifier les métabolites qui expliquent le mieux le phénotype étudié, i.e. faire de la sélection d'effets fixes dans un modèle linéaire mixte. Peu de méthodes existantes répondent à ce problème. La plus performante en grande dimension est le lmmLasso qui est une pénalisation ℓ^1 de la log-vraisemblance du modèle marginal, voir Section 1.4. Cette méthode optimise une fonction objectif non convexe par un algorithme de descente au prix d'une inversion d'une matrice $n \times n$ à chaque itération du processus de convergence, ce qui est relativement coûteux en temps de calcul sur les données métabolomiques. De plus les modèles mixtes sont généralement envisagés avec une seule structure de groupe, c'est-à-dire une seule division des observations, ce qui peut se révéler inapproprié lorsque les observations sont divisées en plusieures structures comme un effet bande et un effet famille où les individus d'une même famille sont répartis dans plusieurs bandes et les bandes contiennent plusieurs familles. La répartition des individus par race et par bande a été donnée dans la Table 2, les statistiques descriptives du nombre d'individus par famille pour les 157 familles que composent les données sont fournies dans la Table 6.

min	1st Q	Median	Mean	3rd Q	Max
1.00	2.00	3.00	3.22	4.00	11.00

TABLE 6 – **Statistiques descriptives sur le nombre d'individus apparentés**

Le package intégrant la méthode lmmLasso (Schelldorfer et al., 2011) n'autorise pas le

cadre des structures chevauchantes, même si le cadre théorique et le modèle (8) considéré ne l'empêchent pas. Nous avons développé une méthode de sélection d'effets fixes qui fonctionne en grande dimension et qui ne nécessite pas d'inversion de matrice $n \times n$, ce qui la rend beaucoup plus rapide que le lmmLasso.
Notre méthode se base sur une autre façon de décrire le modèle linéaire mixte dans laquelle on explicite la matrice V du modèle marginal (8). Considérons un unique effet aléatoire à des fins d'illustrations, alors expliciter la matrice V du modèle (8) conduit au modèle :

$$y = X\beta + Zu + \epsilon, \tag{14}$$

où

- u est un vecteur de taille N correspondant à l'effet aléatoire. On suppose $u \sim \mathcal{N}(0, \sigma_1^2 I_N)$ où σ_1 est un paramètre positif inconnu.
- Z est une matrice d'incidence de taille $n \times N$,
- ϵ est un vecteur gaussien i.i.d. $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I_n)$ où σ_e est un paramètre positif inconnu. On note $R = \sigma_e^2 I_n$.

Il est important de noter que les écritures (8) et (14) sont reliées par l'égalité

$$V = ZGZ' + R.$$

Donnons un exemple simple dans lequel il y a un effet aléatoire qui porte sur l'intercept (comme c'est le cas pour un effet bande ou un effet famille) constitué de 2 groupes ($N = 2$), un exemple de matrice Z est le suivant :

$$Z_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix},$$

où les trois premières observations sont dans le même groupe, et les trois suivantes forment un second groupe.

Notre méthode est une pénalisation ℓ^1 de la vraisemblance complétée, obtenue en considérant les effets aléatoires du modèle (14) comme des données manquantes (comme Bondell et al. (2010) ou Foulley (1997)). La fonction objectif ainsi obtenue est minimisée à l'aide d'un algorithme multicycle ECM (Foulley, 1997; McLachlan and Krishnan, 2008; Meng and Rubin, 1993).
La section suivante présente un article en préparation qui introduit notre nouvelle méthode de sélection d'effets fixes dans un modèle linaire mixte et qui applique cette méthode au jeu de données réelles dont nous disposons.

4.2 Article - Fixed effects selection in high dimensional linear mixed models

Résumé On se place dans le cadre du modèle linéaire mixte dans lequel les observations sont structurées. On propose l'ajout d'une pénalisation ℓ^1 portant sur les effets fixes dans la log-vraisemblance complétée, obtenue en considérant les effets aléatoires comme des données manquantes. Un algorithme 'multicycle ECM' est utilisé pour résoudre le problème d'optimisation ; cet algorithme peut être combiné à n'importe quelle méthode de sélection de variables développée pour le modèle linéaire classique. La méthode proposée fonctionne lorsque le nombre de paramètres p est plus grand que le nombre d'observations n ; elle est plus rapide que le lmmLasso (Schelldorfer et al., 2011) puisque ne nécessitant pas l'inversion d'une matrice de taille $n \times n$ à chaque itération du processus de convergence. Des résultats théoriques sont fournis dans le cas où les variances des effets aléatoires et de la résiduelle sont connues. La combinaison de l'algorithme avec la méthode procbol (Rohart, 2012) donne de très bons résultats sur l'estimation de l'ensemble des effets fixes ainsi que l'estimation des variances ; ces résultats sont meilleurs que ceux du lmmLasso, en petite dimension ($p < n$) mais aussi en grande dimension ($p > n$).

Article soumis

Fixed effects Selection in high dimensional Linear Mixed Models

Florian Rohart, Magali San-Cristobal and Béatrice Laurent

2012

Abstract

We consider linear mixed models in which the observations are grouped. A ℓ^1-penalization on the fixed effects coefficients of the log-likelihood obtained by considering the random effects as missing values is proposed. A multicycle ECM algorithm is used to solve the optimization problem; it can be combined with any variable selection method developed for linear models. The algorithm allows the number of parameters p to be larger than the total number of observations n; it is faster than the lmmLasso (Schelldorfer et al., 2011) since no $n \times n$ matrix has to be inverted. We show that the theoretical results of Schelldorfer et al. (2011) apply for our method when the variances of both the random effects and the residuals are known. The combination of the algorithm with a variable selection method (Rohart, 2011) shows good results in estimating the set of relevant fixed effects coefficients as well as estimating the variances; it outperforms the lmmLasso both in the common case ($p < n$) and in the high-dimensional case ($p \geq n$).

1 Introduction

More and more real data sets are high-dimensional data because of the widely-used new technologies such as high-thoughput DNA/RNA chips or RNA seq in biology. The high-dimensional setting -in which the number of parameters p is greater than the number of observations n- generally implies that the problem can not be solved. In order to address this problem, some conditions are usually added such as a sparsity condition -which means that a lot of parameters are equal to zero- or a well-conditioning of the variance matrix of the observations, among others. A lot of work has been done to address the problem of variable selection, mainly in a linear model $Y = X\beta + \epsilon$, where X is an $n \times p$ matrix containing the observations and ϵ is a n-vector of i.i.d random variables, usually Gaussian. One of the oldest method is the Akaike Information Criterion (AIC), which is a penalization of the log-likelihood by a function of the number of parameters included in the model. More recently, the Lasso (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) revolutionized the field with both a simple and powerful method: ℓ^1-penalization of the least squares estimate which exactly shrinks to zero some coefficients. The Lasso has some

extensions, a group Lasso (Yuan and Lin, 2007), an adaptive Lasso (Huang et al., 2008) and a more stable version known as BoLasso (Bach, 2009), for example. A penalization on the likelihood is not the only way to perform variable selection. Indeed statistical testing has also been used recently (Rohart, 2011) and it appears to give good results.

In all methods cited above, the observations are supposed to be independent and identically distributed. When a structure information is available, such as family relationships or common environmental effects, these methods are no longer adapted. In a linear mixed model, the observations are assumed to be clustered, hence the variance-covariance matrix V of the observations is no longer diagonal but could be assumed to be block diagonal in some cases. A lot of literature about linear mixed models concerns the estimation of the variance components, either with a maximum likelihood estimation (ML) (Henderson, 1973, 1953) or a restricted maximum likelihood estimation (REML) which accounts for the loss in degrees of freedom due to fitting fixed effects (Patterson and Thompson, 1971; Harville, 1977; Henderson, 1984; Foulley et al., 2006). However, both methods assume that each fixed effect and each random effect is relevant. This assumption might be wrong and leads to false estimation of the parameters, especially in a high-dimensional analysis. Contrary to the linear model, there is little literature about selection of fixed effects coefficients in a linear mixed model in a high-dimensional setting.

Both Bondell et al. (2010) and Ibrahim et al. (2011) used a penalized likelihood to perform selection of both the fixed and the random effects. However, their simulation studies were only designed in a low dimensional context. Bondell et al. (2010) introduced a constrained EM algorithm to solve the optimization problem, however the algorithm does not really cope with the problem of high dimension. To our knowledge, only Schelldorfer et al. (2011) studied the topic in a high dimensional setting. Their paper introduced an algorithm based on a ℓ^1-penalization of the maximum likelihood estimator in order to select the relevant fixed effects coefficients. As highlighted in their paper, their algorithm relies on the inversion of the variance matrix of the observations V, which can be time-consuming. Finally, their method depends on a regularization parameter that has to be tuned, as for the original Lasso. As this question remains an open problem, they proposed the use of the Bayesian Information Criterion (BIC) to choose the penalty.
All methods are usually considered with one grouping factor -meaning one partition of the observations-, which can be sometimes misappropriate when the observations are divided w.r.t two factors or more; for instance when a family relationship and a common environmental effect are considered.

We present in this paper another way to perform selection of the fixed effects in a linear mixed model. We propose to consider the random effects as missing data, as done in Bondell et al. (2010) or in Foulley (1997), and to add a ℓ^1-penalization on the log-likelihood of the complete data. Our method allows the use of several different grouping factors. We propose a multicycle ECM algorithm (Foulley, 1997; McLachlan and Krishnan,

2008; Meng and Rubin, 1993) to solve the optimization problem; this algorithm possesses convergence properties. In addition, we show that the use of BIC in order to tune the regularization parameter as proposed by Schelldorfer et al. (2011) could sometimes turn out to be misappropriate.
We give theoretical results when the variances of the observations are known. Due to the design of the algorithm that is decomposed into steps, the algorithm can be combined with any variable selection method built for linear models. Nevertheless, the performance of the combination strongly depends on the variable selection method that is used. As there is little literature on the selection of the fixed effects in a high-dimensional linear mixed model, we will mainly compare our results to those of Schelldorfer et al. (2011).

This paper extends the analysis on a real data-set coming from a project in which hundreds of pigs have been studied. The aim is to enlighten relationships between some phenotypes of interest and metabolomic data (Rohart et al., 2012). Linear mixed models are appropriate since the observations are repeated data from different environments (groups of animals are reared together in the same conditions). Some individuals are also genetically related, in a family effect. The data set consists in 506 individuals from 3 breeds, 8 environments and 157 families. The metabolomic data contains $p = 375$ variables. We will investigate the Daily Feed Intake (DFI) phenotype.

This paper is organized as follows: we will first describe the linear mixed model and the objective function, then we will present the multicycle ECM algorithm that is used to solve the optimization problem of the objective function. Section 3 gives a generalization of the algorithm of Section 2 that can be used with any variable selection method developed for linear models. Finally, we will present results from a simulation study showing that the combination of this new algorithm with a good variable selection method performs well, in terms of selection of both the fixed and random effects coefficients (Section 4), before applying the method on a real data set in Section 5.

2 The method

Let us introduce some notations that will be used throughout the paper. $Var(a)$ denotes the variance-covariance matrix of the vector a. For all $a > 0$, set I_a to be the identity matrix of $\mathbb{R}^a$. For $A \in \mathbb{R}^{n\times p}$, let $A_{I,J}$ $A_{.,J}$ and $A_{I,.}$ denote respectively the submatrix of A composed of elements of A whose rows are in I and columns are in J, whose columns are in J with all rows, and whose rows are in I with all columns. Moreover, we set for all $a > 0, b > 0$, 0_a to be the vector of size a with all its coordinates equal to 0 and $0_{a\times b}$ to be the null matrix of size $a \times b$. Let us denote $|A|$ the determinant of matrix A.

2.1 The linear mixed model setup

We consider the linear mixed model in which the observations are grouped and we suppose that only a small subset of the fixed effects coefficients are non-zero. The aim of this paper is to recover this subset through an algorithm that will be presented in the next section. In the present section we explicit the linear mixed model and our objective function.

Mixed models are often considered with a single grouping factor, meaning that each observation belongs to one single group. In this paper we allow several grouping factors. Assume there are q random effects and q grouping factors ($q \geq 1$), where some grouping factors may be identical. The levels of the factor k are denoted $\{1, 2, \ldots, N_k\}$. The i^{th}-observation belongs to the groups $(i_1, \ldots, i_q)$, where for all $l = 1, \ldots, q$, $i_l \in \{1, 2, \ldots, N_l\}$. We precise that two observations can belong to the same group of one grouping factor whereas they can belong to different groups of another grouping factor.

Let n be the total number of observations with $n = \sum_{i=1}^{N_k} n_{i,k}, \forall k \leq q$, where $n_{i,k}$ is the number of observations within group i from the grouping factor k. Denote $N = \sum_{k=1}^{q} N_k$.

The linear mixed model can be written as

$$y = X\beta + \sum_{k=1}^{q} Z_k u_k + \epsilon, \tag{1}$$

where

- y is the set of observed data of length n,
- β is an unknown vector of $\mathbb{R}^p$; $\beta = (\beta_1, \ldots, \beta_p)$,
- X is the $n \times p$ matrix of fixed effects; $X = (X_1, \ldots, X_p)$,
- For $k = 1, \ldots, q$, u_k is a N_k-vector of the random effect corresponding to the grouping factor k, ,
- For $k = 1, \ldots, q$, Z_k is a $n \times N_k$ incidence matrix corresponding to the grouping factor k,
- $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ is a Gaussian vector with i.i.d. components $\epsilon \sim \mathcal{N}_n(0, \sigma_e^2 I_n)$, where σ_e is an unknown positive quantity. We denote by R the variance-covariance matrix of ϵ, $R = \sigma_e^2 I_n$.

To fix ideas, let us give a example of matrices Z_k for $n = 6$ and two random effects.

Let $Z_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ and $Z_2 = \begin{pmatrix} x_1 & 0 & 0 \\ x_2 & 0 & 0 \\ 0 & x_3 & 0 \\ 0 & x_4 & 0 \\ 0 & 0 & x_5 \\ 0 & 0 & x_6 \end{pmatrix}$. The grouping factors 1 and 2 are the same for the two random effects u_1 and u_2, and Z_2 is the incidence matrix of the interaction of the variable $x = (x_1, \ldots, x_6)$ and the grouping factor.

Throughout the paper, we assume that $u_k \sim \mathcal{N}_{N_k}(0, \sigma_k^2 I_{N_k})$, where σ_k is an unknown positive quantity. We denote $u = (u_1', \ldots, u_k')'$, Z the concatenation of $(Z_1, \ldots, Z_q)$, G the block diagonal matrix of $\sigma_1^2 I_{N_1}, \ldots, \sigma_q^2 I_{N_q}$ and Γ the block diagonal matrix of $\gamma_1 I_{N_1}, \ldots, \gamma_q I_{N_q}$, where $\gamma_k = \sigma_e^2/\sigma_k^2$.
Remark that with these notations, Model (1) can also be written as: $y = X\beta + Zu + \epsilon$.
In the following, we assume that $\epsilon, u_1, \ldots, u_q$ are mutually independent. Thus $Var(u_1, \ldots, u_q, \epsilon) = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}$. We consider the matrices X and $\{Z_k\}_{1,\ldots,q}$ to be fixed design.

Note that our model (1) and the one in Schelldorfer et al. (2011) are almost identical when all the grouping factors are identical, except that we supposed $u_1 \ldots, u_q$ to be independent while they did not make this assumption. Nevertheless, for their simulation study, they considered i.i.d. random effects.

Let us denote by J the set of the indices of the relevant fixed effects of Model (1); $J = \{j, \beta_j \neq 0\}$. The aim of this paper is to estimate J, β, G and R. In the whole paper, the number of fixed effects p can be larger than the total number of observations n. However, we focus on the case where only a few fixed-effects are relevant. We also assume that only a few grouping factors are included in the model since this paper was motivated by such a case on a real data set, see Section 5. Hence we assume $N + |J| < n$.

2.2 A ℓ^1 penalization of the complete log-likelihood

In the following, we consider the fixed effects coefficients β and the variances $\sigma_1^2, \ldots, \sigma_q^2, \sigma_e^2$ as parameters and $\{u_k\}_{k\in\{1,\ldots,q\}}$ as missing data. We denote $\Phi = (\beta, \sigma_1^2, \ldots, \sigma_q^2, \sigma_e^2)$.
The log-likelihood of the complete data $x = (y, u)$ is

$$L(\Phi; x) = L_0(\beta, \sigma_e^2, \sigma_1^2, \ldots, \sigma_q^2; \epsilon) + \sum_{k=1}^{q} L_k(\sigma_k^2; u_k), \tag{2}$$

where

$$-2L_0(\beta, \sigma_e^2, \sigma_1^2, \ldots, \sigma_q^2; \epsilon) = n\log(2\pi) + n\log(\sigma_e^2) + \left|\left| y - X\beta - \sum_{k=1}^{q} Z_k u_k \right|\right|^2 / \sigma_e^2, \quad (3a)$$

$$\forall k \in \{1, \ldots, q\}, -2L_k(\sigma_k^2; u_k) = N_k \log(2\pi) + N_k \log(\sigma_k^2) + ||u_k||^2 / \sigma_k^2. \quad (3b)$$

Indeed, (2) comes from $p(x|\Phi) = p(y|\beta, u_1, \ldots, u_q, \sigma_e^2)\Pi_{k=1}^q p(u|\sigma_k^2)$; (3a) comes from $L_0(\beta, \sigma_e^2, \sigma_1^2, \ldots, \sigma_q^2; \epsilon) = L_0(\sigma_e^2; \epsilon) = n\log(2\pi) + n\log(\sigma_e^2) + \epsilon'\epsilon/\sigma_e^2$ because $\epsilon|\sigma_e^2 \sim \mathcal{N}_n(0, \sigma_e^2 I_n)$ and (3b) from $u_k|\sigma_k^2 \sim \mathcal{N}_{N_k}(0, \sigma_k^2 I_{N_k})$.

Since we allow the number of fixed-effects p to be larger than the total number of observations n, the usual maximum likelihood (ML) or restricted maximum likelihood (REML) approaches do not apply. As we assumed that β is sparse -many coefficients are assumed to be null- and since we want to recover that sparsity, we add a ℓ^1 penalty on β to the log-likelihood of the complete data (2). Indeed a ℓ^1 penalization is known to induce sparsity in the solution, as in the Lasso method (Tibshirani, 1996) or the lmmLasso method (Schelldorfer et al., 2011). Thus we consider the following objective function to be minimized:

$$g(\Phi; x) = -2L(\Phi; x) + \lambda|\beta|_1, \quad (4)$$

where λ is a positive regularization parameter. Remark that the function g could have been obtained from a Bayesian setting considering a Laplace prior on β.
It is interesting to note that finding a minimum of the objective function (4) is a non-linear, non-differentiable and non convex problem. But more importantly, one thing that strikes out -especially from (3b)- is that the function g is not lower-bounded. Indeed, $L(\Phi; x)$ tends to infinity when both u_k and σ_k tends toward 0. It is a well-known problem of degeneracy of the likelihood, especially studied in Gaussian mixture model (Biernacki and Chrétien, 2003) but not much concerning mixed models. In linear mixed models, some authors focus on the log-likelihood of the marginal model in which the random effects are integrated out in the matrix of variance of the observations Y, such as in Schelldorfer et al. (2011):

$$y = X\beta + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, V).$$

Note that $V = ZGZ' + R$. The degeneracy of the likelihood can also appear in the marginal model when the determinant of V tends toward zero. This phenomenon is likely to happen in a high dimensional context when too much fixed-effects enter the model, that is to say when the amount of regularization chosen by the penalty of the lmmLasso (Schelldorfer et al., 2011) or by λ in (4) is not large enough.

Because of the non lower-boundness of the likelihood, the problem of minimizing the function g is ill-posed: we are not interested in the minimization of g on the parameter space $\left\{\beta \in \mathbb{R}^p, \sigma_1^2 \geq 0, \ldots, \sigma_q^2 \geq 0, \sigma_e^2 \geq 0\right\}$ but more interested in minimizing g inside the

parameter space

$$\Lambda = \left\{\beta \in \mathbb{R}^p, \sigma_1^2 > 0, \ldots, \sigma_q^2 > 0, \sigma_e^2 > 0\right\}.$$

Instead of adding a ℓ^1 penalty on the random effect as Bondell et al. (2010), we will use the degeneracy of the likelihood at the frontier of the parameter space Λ to perform selection of the random effects. Indeed, if it exists $1 \leq k \leq q$ such that the minimization process of the function g, defined by (4), takes place at the frontier $\sigma_k^2 = 0$ of the parameter space Λ, then the grouping factor k is deleted from the model (1). Nevertheless, our method is more restrictive than the one of Bondell et al. (2010) since we assume $N+|J| < n$.

The minimization process of the function g can coincide with the deletion of the random effect k, for $1 \leq k \leq q$, for two reasons: either the true underlying model was different from the fitted one -some grouping factors are included in the model although there is no need to-, or because the initialization of the minimization process was to close to an attraction domain of $(u_k, \sigma_k^2) = (0_{N_k}, 0)$ (Biernacki and Chrétien, 2003).

When selection of the random effects is performed in the linear mixed model (1) with q random effects, a new model is fitted with $q-1$ grouping factor and the objective function is modified accordingly. The selection of the random effects can be performed until no grouping factor remains, then a linear model is considered.

In the next section we will use a multicycle ECM algorithm in order to solve the minimization of (4); it performs selection of both the fixed and the random effects.

2.3 A multicycle ECM algorithm

The multicycle ECM algorithm (Meng and Rubin, 1993; Foulley, 1997; McLachlan and Krishnan, 2008) used to solve the minimization problem of (4) contains four steps -two E steps interlaced with two M steps-; each will be described in this section.

Recall that $\Phi = (\beta, \sigma_1^2, \ldots, \sigma_q^2, \sigma_e^2)$ is the vector of the parameters to estimate and that $u = (u_1', \ldots, u_k')'$ is a vector of missing values. For the sake of simplicity, we denote $\mathcal{K} = \{1, \ldots, q\}$ and $\sigma_{\mathcal{K}}^2 = \{\sigma_k^2\}_{k \in \mathcal{K}}$.

The multicyle ECM algorithm is an iterative algorithm. We will index the iterations by $t \in \mathbb{N}$. $\Theta^{[t]}$ will denote the current estimation of the parameter Θ at iteration t.

Let $E_{u|y,\Phi=\Phi^{[t]}}$ denote the conditional expectation under the distribution of u given the vector of observations y and the current estimation of the set of parameters Φ at iteration t.

2.3.1 First E-step

Let denote

$$Q(\Phi; \Phi^{[t]}) = E_{u|y,\Phi=\Phi^{[t]}}[g(\Phi; x)].$$

We can decompose Q as follows:

$$Q(\Phi;\Phi^{[t]}) = Q_0(\beta,\sigma^2_{\mathcal{K}},\sigma^2_e;\Phi^{[t]}) + \sum_{k=1}^{q} Q_k(\sigma^2_k;\Phi^{[t]}),$$

where

$$Q_0(\Phi;\Phi^{[t]}) = n\ \log(2\pi) + n\ \log(\sigma_e^{2[t]}) + E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon)/\sigma_e^{2[t]} + \lambda|\beta^{[t]}|_1$$

and

$$\forall k \in \mathcal{K}, Q_k(\sigma^2_k;\Phi^{[t]}) = N\log(2\pi) + N\log(\sigma_k^{2[t]}) + E_{u|y,\Phi=\Phi^{[t]}}({u_k}'u_k)/\sigma_k^{2[t]}.$$

By definition, we have for all $1 \leq i \leq n, Var_{u|y,\Phi=\Phi^{(t)}}(\epsilon_i) = E_{u|y,\Phi=\Phi^{[t]}}(\epsilon_i^2) - \left|\left|E_{u|y,\Phi=\Phi^{[t]}}(\epsilon_i)\right|\right|^2$. Hence

$$E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon) = \left|\left|E_{u|y,\Phi=\Phi^{[t]}}(\epsilon)\right|\right|^2 + tr\left(Var_{u|y,\Phi=\Phi^{(t)}}(\epsilon)\right).$$

We can then explicit

$$E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon) = \left|\left|y - X\beta^{[t]} - ZE\left(u|y,\Phi=\Phi^{[t]}\right)\right|\right|^2 + tr\left(ZVar\left(u|y,\Phi^{[t]}\right)Z'\right). \quad (5)$$

According to the denomination of Henderson (1973), $E\left(u|y,\Phi=\Phi^{[t]}\right)$ is the BLUP (Best Linear Unbiased Prediction) of u for the vector of parameters Φ equal to $\Phi^{[t]}$. Let us denote $u^{[t+1/2]} = E\left(u|y,\Phi=\Phi^{[t]}\right)$, we have that

$$u^{[t+1/2]} = (Z'Z + \Gamma^{[t]})^{-1}Z'\left(y - X\beta^{[t]}\right).$$

2.3.2 M-Step for β

The next step performs a minimization of $Q_0(\beta,\sigma^2_{\mathcal{K}},\sigma^2_e;\Phi^{[t]})$ with respect to β:

$$\beta^{[t+1]} = \underset{\beta}{Argmin}\left(\frac{1}{\sigma_e^{2[t]}}\left|\left|\left(y - Zu^{[t+1/2]}\right) - X\beta\right|\right|^2 + \lambda\left|\beta\right|_1\right). \quad (6)$$

Remark that (6) is a Lasso on β with the vector of "observed" data $\left(y - Zu^{[t+1/2]}\right)$ and the penalty $\lambda\sigma_e^{2[t]}$.

2.3.3 Second E-Step

A second E-step is performed with the actualization of the vector of missing values u: $u^{[t+1]} = E\left(u|y,\beta=\beta^{[t+1]},\sigma_1^2=\sigma_1^{2[t]},\ldots,\sigma_q^2=\sigma_q^{2[t]},\sigma_e^2=\sigma_e^{2[t]}\right)$, thus

$$u^{[t+1]} = (Z'Z + \Gamma^{[t]})^{-1}Z'\left(y - X\beta^{[t+1]}\right).$$

We define $\forall k \in \mathcal{K}$, $u_k^{[t+1]}$ to be the element of size N_k that corresponds to the grouping factor k in $u^{[t+1]}$.

2.3.4 M-step for $(\sigma_1^2, \ldots, \sigma_q^2, \sigma_e^2)$

The actualization of the variances $\{\sigma_k^2\}_{1\leq k\leq q}$ and σ_e^2 are performed with the minimization of $\{Q_k\}_{1\leq k\leq q}$ and Q_0 respectively.
Let $k \in \mathcal{K}$, the minimization of Q_k with respect to σ_k^2 gives:
$\sigma_k^{2[t+1]} = E\left(u_k' u_k | y, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right)/N_k$.
Besides,

$$E\left(u_k' u_k | y, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right) = \left|\left|E\left(u_k | y, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right)\right|\right|^2 + tr\left(Var\left(u_k | y, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right)\right).$$

Moreover we have, thanks to Henderson (1973),

$$Var\left(u_k | y, \sigma_k^{2[t]}, \sigma_e^{2[t]}, \beta^{[t+1]}\right) = T_{k,k}\sigma_e^{2[t]},$$

where $T_{k,k}$ is defined as follows:

$$\begin{aligned}\left(Z'Z + \Gamma^{[t]}\right)^{-1} &= \begin{pmatrix} Z_1'Z_1 + \gamma_1^{[t]} I_{N_1} & Z_1'Z_2 & \ldots & Z_1'Z_q \\ Z_2'Z_1 & Z_2'Z_2 + \gamma_2^{[t]} I_{N_2} & \ldots & Z_2'Z_q \\ \vdots & \vdots & \ddots & \vdots \\ Z_q'Z_1 & Z_q'Z_2 & \ldots & Z_q'Z_q + \gamma_q^{[t]} I_{N_q} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} T_{1,1} & T_{1,2} & \ldots & T_{1,q} \\ T_{1,2}' & T_{2,2} & \ldots & T_{2,q} \\ \vdots & \vdots & \ddots & \vdots \\ T_{1,q}' & T_{2,q}' & \ldots & T_{q,q} \end{pmatrix}.\end{aligned}$$

Thus, for all $k \in \mathcal{K}$:

$$\sigma_k^{2[t+1]} = \frac{1}{N_k}\left[\left|\left|u_k^{[t+1]}\right|\right|^2 + tr\left(T_{k,k}\right)\sigma_e^{2[t]}\right].$$

The minimization of Q_0 with respect to σ_e^2 gives: $\sigma_e^{2[t+1]} = E_{u|y,\Phi=\Phi^{[t]}}(\epsilon'\epsilon)/n$. From (5), we have

$$\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Zu^{[t+1]}\right|\right|^2 + tr\left(Z(Z'Z + \Gamma^{[t]})^{-1}Z'\right)\sigma_e^{2[t]}\right].$$

Since

$$\begin{aligned} tr\left(Z\left(Z'Z+\Gamma^{(t)}\right)^{-1}Z'\right) &= tr\left(\left(Z'Z+\Gamma^{(t)}\right)^{-1}Z'Z\right) \\ &= N - tr\left[\left(Z'Z+\Gamma^{(t)}\right)^{-1}\Gamma^{[t]}\right] \\ &= N - \sum_{k=1}^{q}\gamma_k^{[t]}tr\left(T_{k,k}\right) \end{aligned}$$

we have

$$\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Zu^{[t+1]}\right|\right|^2 + \left(N - \sum_{k=1}^{q}\gamma_k^{[t]}tr\left(T_{k,k}\right)\right)\sigma_e^{2[t]}\right].$$

In summary, the algorithm is the following:

Algorithm 2.1 (Lasso+)**.** *Initialization:*
Set $\mathcal{K} = \{1, \ldots, q\}$. *Initialize the set of parameters* $\Phi^{[0]} = (\sigma_{\mathcal{K}}^{2[0]}, \sigma_e^{2[0]}, \beta^{[0]})$.
Define $\Gamma^{[0]}$ *as the block diagonal matrix of* $\gamma_1^{[0]}I_{N_1}, \ldots, \gamma_q^{[0]}I_{N_q}$, *where* $\gamma_k^{[0]} = \sigma_e^{2[0]}/\sigma_k^{2[0]}$.
Define Z *as the concatenation of* $Z_1, \ldots, Z_q$ *and* $u = (u_1', \ldots, u_q')'$.
Until convergence:
1. E-step
$u^{[t+1/2]} = (Z'Z + \Gamma^{[t]})^{-1}Z'(y - X\beta^{[t]})$
2. M-step
$\beta^{[t+1]} = \underset{\beta}{Argmin}\left(\left|\left|\left(y - Zu^{[t+1/2]}\right) - X\beta\right|\right|^2 + \lambda\sigma_e^{2[t]}\left|\beta\right|_1\right)$
3. E-step
$u^{[t+1]} = (Z'Z + \Gamma^{[t]})^{-1}Z'(y - X\beta^{[t+1]})$
4. M-step
(a) For k *in* $\mathcal{K}$, *set* $\sigma_k^{2[t+1]} = \left|\left|u_k^{[t+1]}\right|\right|^2/N_k + tr\left(T_{k,k}\right)\sigma_e^{2[t]}/N_k$
(b) Set $\sigma_e^{2[t+1]} = \frac{1}{n}\left[\left|\left|y - X\beta^{[t+1]} - Zu^{[t+1]}\right|\right|^2 + \sum_{k\in\mathcal{K}}\left(N_k - \gamma_k^{[t]}tr\left(T_{k,k}\right)\right)\sigma_e^{2[t]}\right]$
(c) For k *in* $\mathcal{K}$, *if* $\left(\left|\left|u_k^{[t+1]}\right|\right|^2/N_k < 10^{-4}\sigma_e^{2[t]}\right)$ *then* $\mathcal{K} = \mathcal{K}\backslash\{k\}$
Define Z *as the concatenation of* $\{Z_k\}_{k\in\mathcal{K}}$ *and* u *as the transpose of the concatenation of* $\{u_k'\}_{k\in\mathcal{K}}$.
Set $\Gamma^{[t+1]}$ *as the block diagonal matrix of* $\left\{\gamma_k^{[t+1]}I_{N_k}\right\}_{k\in\mathcal{K}}$, *where for all* $k \in \mathcal{K}$,
$\gamma_k^{[t+1]} = \sigma_e^{2[t+1]}/\sigma_k^{2[t+1]}$.
end

The convergence of Algorithm 2.1 is ensured since it is a multicycle ECM algorithm (Meng and Rubin, 1993).
Three stopping criteria are used to stop the convergence process of the algorithm: a condition on $||\beta^{[t+1]} - \beta^{[t]}||^2$, a condition on $||u_k^{[t+1]} - u_k^{[t]}||^2$ for each random effect u_k and

a condition on $||L(\Phi^{[t+1]}, x) - L(\Phi^{[t]}, x)||^2$ where $L(\Phi, x)$ is the log-likelihood defined by (2). The convergence takes place when all the criteria are fulfilled. We also add a fourth condition that controls the number of iterations. We choose to initialize the algorithm 2.1 as follows: for all $1 \leq k \leq q, \sigma_k^{2[0]} = \frac{0.4}{q}\ \sigma_e^{2[-1]}, \sigma_e^{2[0]} = 0.6\ \sigma_e^{2[-1]}$, and $(\sigma_e^{2[-1]}, \beta^{[0]})$ is estimated from a linear estimation (without the random effects) of the Lasso at the given penalty λ. We will study in Section 4.4 the influence of the initialization of the algorithm on simulated data.

Note that Step 4(c) performs the selection on the random effects; we decide to delete a random effect when its variance became lower that $10^{-4}\sigma_e^{2[t]}$.

The estimation of the set of parameters Φ is biased (Zhang and Hunag, 2008). One last step can be added in order to address this problem once both Algorithm 2.1 has converged and the penalization parameter λ has be tuned. Indeed, one should prefer to use Algorithm 2.1 in order to estimate both the support of β and the support of the random effects, and then to estimate the set Φ with a classical mixed model estimation on the model:

$$y = X\beta_{\hat{J}} + \sum_{k \in S} Z_k u_k + \epsilon,$$

where $\hat{J}$ and S are the estimated set of indices of the relevant fixed effects and the estimated set of indices of the relevant random effects respectively.

Proposition 2.2. *When the variances are known, the minimization of our objective function* (4) *is the same as the minimization of* $Q(\beta) = (y - X\beta)'V^{-1}(y - X\beta) + \lambda|\beta|_1$, *which is the objective function of Schelldorfer et al. (2011) at known variances.*

Let us recall that Schelldorfer et al. (2011) obtained theoretical results on the consistency of their method. According to Proposition 2.2, these results apply to our method in the case of known variances. The proof of Proposition 2.2 is given in Web Appendix C.

Note that when individuals are genetically related through a known relationship matrix A, we have $u \sim \mathcal{N}_n(0, \sigma_s^2 A)$, with $\sigma_s > 0$. Thanks to Henderson (1973), A^{-1} can be directly computed. In all that precede, the changes are the following : the matrix Γ becomes the matrix $\sigma_e^2/\sigma_s^2 A^{-1}$ and $||u||^2$ becomes $u'A^{-1}u$.

2.4 The tuning parameter

Algorithms 2.1 involves a regularization parameter λ; the solution depends on this parameter. This amount of shrinkage has to be tuned. We choose the use of the Bayesian

Information Criterion (BIC) (Schwarz, 1978):

$$\lambda_{BIC} = \underset{\lambda}{Argmin} \left\{ \log|V_\lambda| + (y - X\hat{\beta}_\lambda)' V_\lambda^{-1} (y - X\hat{\beta}_\lambda) + d_\lambda . \log(n) \right\},$$

where $V_\lambda = \sum_{k\in\mathcal{K}} \hat{\sigma}_k^2 Z_k Z_k' + \hat{\sigma}_e^2 I_n$ and $\hat{\sigma}_k^2, \hat{\sigma}_e^2, \hat{\beta}_\lambda$ are obtained from the minimization of the objective function g defined by (4). Moreover, $d_\lambda := \sum_{k=1}^{p} 1_{\sigma_k \neq 0} + |\hat{J}_\lambda|$ is the sum of the number of non-zero variance-covariance parameters and the number of non-zero fixed effects coefficients included in the model which has been selected with the regularization parameter λ.

Other methods can be used to choose λ such as AIC or cross-validation, among others. An advantage of BIC over cross-validation is mainly the gain of computational time.

In the next section, we propose a generalization of Algorithm 2.1 which allows the use of any variable selection methods developed for linear models.

3 A generalized algorithm

Algorithm 2.1 gives good results, as it can be seen in the simulation study of Section 4. Nevertheless, since Step 2 of Algorithm 2.1 aims at selecting the relevant coefficients of β in a linear model, the Lasso method can be replaced with any variable selection method built for linear models. If the chosen variable selection method optimizes a criterion, such as the adaptive Lasso (Zou, 2006) or the elastic net (Zou and Hastie, 2005), the algorithm thus obtained remains a multicycle ECM algorithm and the convergence property still applies. However, the convergence property does not hold for methods that do not optimize a criterion.

Algorithm 2.1 can be reshaped for a generalized algorithm as follows:

Algorithm 3.1. *Initialization:*
Initialize the set of parameters $\Phi^{[0]} = (\sigma_{\mathcal{K}}^{2[0]}, \sigma_e^{2[0]}, \beta^{[0]})$. *Set* $\mathcal{K} = \{1, \ldots, q\}$.
Define $\Gamma^{[0]}$ *as the block diagonal matrix of* $\gamma_1^{[0]} I_{N_1}, \ldots, \gamma_q^{[0]} I_{N_q}$, *where* $\gamma_k^{[0]} = \sigma_e^{2[0]}/\sigma_k^{2[0]}$.
Define Z *as the concatenation of* $Z_1, \ldots, Z_q$ *and* $u = (u_1', \ldots, u_q')'$.
Until convergence:
1. $u^{[t+1/2]} = (Z'Z + \Gamma^{[t]})^{-1} Z'(y - X\beta^{[t]})$
2. Variable selection and estimation of β *in the linear model* $y - Zu^{[t+1/2]} = X\beta + \epsilon^{[t]}$, *where* $\epsilon^{[t]} \sim \mathcal{N}(0, \sigma_e^{2[t]} I_n)$.
3. $u^{[t+1]} = (Z'Z + \Gamma^{[t]})^{-1} Z'(y - X\beta^{[t+1]})$
4. (a) For k *in* $\mathcal{K}$, *set* $\sigma_k^{2[t+1]} = \left|\left| u_k^{[t+1]} \right|\right|^2 / N_k + tr\left(T_{k,k}\right) \sigma_e^{2[t]} / N_k$
(b) Set $\sigma_e^{2(t+1)} = \frac{1}{n} \left[\left|\left| y - X\beta^{[t+1]} - Zu^{[t+1]} \right|\right|^2 + \sum_{k\in\mathcal{K}} \left(N_k - \gamma_k^{[t]} tr\left(T_{k,k}\right) \right) \sigma_e^{2[t]} \right]$
(c) For k *in* $\mathcal{K}$, *if* $\left(\left|\left| u_k^{[t+1]} \right|\right|^2 / N_k < 10^{-4} \sigma_e^{2[t]} \right)$ *then* $\mathcal{K} = \mathcal{K} \backslash \{k\}$

Define Z as the concatenation of $\{Z_k\}_{k\in\mathcal{K}}$ and u as the transpose of the concatenation of $\{u'_k\}_{k\in\mathcal{K}}$.
Set $\Gamma^{[t+1]}$ as the block diagonal matrix of $\left\{\gamma_k^{[t+1]} I_{N_k}\right\}_{k\in\mathcal{K}}$, where for all $k \in \mathcal{K}$,
$\gamma_k^{[t+1]} = \sigma_e^{2[t+1]}/\sigma_k^{2[t+1]}$.
end

We choose to initialize Algorithm 3.1 as follows: for all $1 \leq k \leq q, \sigma_k^{2[0]} = \frac{0.4}{q}\,\sigma_e^{2[-1]}, \sigma_e^{2[0]} = 0.6\ \sigma_e^{2[-1]}$, and $(\sigma_e^{2[-1]}, \beta^{[0]})$ is estimated from a linear estimation (without the random effects) of the method used at Step 2.
In the following we propose to combine Algorithm 2.1 with a method that does not need a tuning parameter, namely the procbol method (Rohart, 2011). The procbol method is a sequential multiple hypotheses testing which statistically determines the set of relevant variables in a linear model $y = X\beta + \epsilon$ where ϵ is an i.i.d Gaussian noise. This method is a two-step procedure: the first step orders the variables taking into account the observations y and the second step uses multiple hypotheses testing to separate the relevant variables from the irrelevant ones. The procbol method is proved to be powerful under some conditions on the signal in Rohart (2011).

In Section 4, we show that the combination of Algorithm 3.1 and the procbol method performs well on simulated data.

4 Simulation study

The purpose of this section is to compare different methods that aim at selecting both the correct fixed effects coefficients and the relevant random effects in a linear mixed model (1), but also to look at the improvement obtained from including random effects in the model.

4.1 Presentation of the methods

We compare several methods, some of them are designed to work in a linear model: *Lasso* (Tibshirani, 1996), *adLasso* (Zou, 2006) and *procbol* (Rohart, 2011), while others are designed to work in a linear mixed model: *lmmLasso* (Schelldorfer et al., 2011), *Algorithm 2.1* (labelled as *Lasso+*), *adLasso+Algorithm 3.1* (labelled as *adLasso+*) and *procbol+Algorithm 3.1* (labelled as *pbol+*).
The initial weights of the *adLasso* and *adLasso+* are set to be equal to $1/|\hat{\beta}_i|$ where for all $i \in \{1, \ldots, p\}$, $\hat{\beta}_i$ is the Ordinary Least Squares (OLS) estimate of β_i in the model $y = X_i\beta_i + \epsilon_i$.

The second step of the procbol method performs multiple hypotheses testing with an estimation of unknown quantiles related to the matrix X. The calculation of these quantiles

at each iteration of the convergence process would make the combination of the procbol method and Algorithm 3.1 almost impossible to run; however, since the data matrix X stays the same throughout the algorithm, the quantiles also do. Thus the procbol method was adapted to be run several times on the same data set by keeping the calculated quantiles, which led to a enormous gain of computational time. Some parameters of the procbol method were changed in order to limit the time of one iteration of the convergence process, as follows. The parameter m which stands for the number of bootstrapped samples used to sort the variables (first step of the procbol method) was set to 10. The number of variables ordered at the first step of the procbol method was set to 40. Note that when the procbol method was used in a linear model, we set $m = 100$ as advised in Rohart (2011). Both the *procbol* method and the *pbol+* method were set with a user-level of $\alpha \in \{0.1, 0.05\}$, which stands for the level of the testing procedure.
Concerning all methods that needed a tuning parameter, we set it using the Bayesian Information Criterion described in Section 2.4. A particular attention has to be drawn on the tuning of the regularization parameter of some methods that could be tricky in some cases due to the degeneracy of the likelihood, especially *Lasso* and *adLasso*, see Web Appendix B.

4.2 Design of our simulation study

Concerning the design of our simulations, we set X_1 to be the vector of $\mathbb{R}^n$ whose coordinates are all equal to 1 and we considered four models. For each model, the response variable y is computed via $y = \sum_{j=1}^{5} X_{i_j}\beta_{i_j} + \sum_{k=1}^{q} Z_k u_k + \epsilon$, where $J = \{i_1, \ldots, i_5\} \subset \{1, \ldots, p\}$, with two random effects ($q = 2$) being standard Gaussian ($\sigma_1^2 = \sigma_2^2 = 1$) and ϵ being a vector of independent standard Gaussian variables. The models used to fit the data differ in the number of parameters p, the number of random effects q and the dependence structure of the X_i's. For each model, we have that for all $j = 2, \ldots, p$: $\sum_{i=1}^{n} X_{j,i} = 0$ and $\frac{1}{n}\sum_{i=1}^{n} X_{j,i}^2 = 1$. For $k = 1, \ldots, q$, the random effects regression matrix Z_k corresponds to the design matrix of the interaction between the k^{th} column of X and the grouping factor k, which gives a $n \times N_k$ matrix. The design of the matrices Z_k's means that the first q grouping variables generates both a fixed effect (corresponds to β_k's) and a random effect (corresponds to u_k's). As advised in Schelldorfer et al. (2011), the variables that generate both a fixed and a random effect do not undergo feature selection; otherwise the fixed effect coefficients of those variables tends to be shrunken towards 0. The set of variables that do not undergo feature selection can change at each step of the convergence process of our algorithms. Indeed, as soon as a variable does not generate a random effect anymore, the fixed effect corresponding to that variable undergoes feature selection again.
The models are defined as follows:

- M_1: $n = 120$, $p = 80$, $\beta_J = 2/3$. For all $j = 2, \ldots, p, X_j \sim \mathcal{N}_n(0, I_n)$.The division

of the observations for the two random effects are the same; for all $k \leq 2 : N_k = 20, \forall i \in \{1, .., 20\}\ n_{i,k} = 6$. This model is fitted assuming $q = 3$.

- M_2: $n = 120$, $p = 300$, $\beta_J = 3/4$. The covariates are generated from a multivariate normal distribution with mean zero and covariance matrix Σ with the pairwise correlation $\Sigma_{kk'} = \rho^{|k-k'|}$ and $\rho = 0.5$. The division of the observations for the two random effects are the same; for all $k \leq 2 : N_k = 20, \forall i \in \{1, .., 20\}\ n_{i,k} = 6$.

- M_3: $n = 120$, $p = 300$, $\beta_J = 2/3$. For all $j = 2, \ldots, p, X_j \sim \mathcal{N}_n(0, I_n)$. The division of the observations for the two random effects are different: $N_1 = 20, \forall i \in \{1, .., 20\}\ n_{i,1} = 6$ and $N_2 = 15, \forall i \in \{1, .., 15\}\ n_{i,2} = 8$

- M_4: $n = 120$, $p = 600$, $\beta_J = 2/3$. For all $j = 2, \ldots, p, X_j \sim \mathcal{N}_n(0, I_n)$. The division of the observations for the two random effects are the same; for all $k \leq 2 : N_k = 20, \forall i \in \{1, .., 20\}\ n_{i,k} = 6$.

For models M_1, M_3, M_4, we set $J = \{1, \ldots, 5\}$. For model M_2, we set $J = \{1, 2, i_3, i_4, i_5\}$ where $\{i_3, i_4, i_5\} \subset \{3, \ldots, p\}$.

In each model, the aim is to recover both the set of relevant fixed effects coefficients J and the set of relevant random effects; but also to estimate the variance of both the random effects and the residuals. To judge the quality of the methods, we use several criterion: the percentage of true model recovered under the label 'Truth' (both J and the set of relevant random effects), the percentage of times the true set of fixed effects is recovered '$\hat{J} = J$', the cardinal of the estimated set of fixed effects coefficients $|\hat{J}|$, the number of true positive TP, the estimated variance $\hat{\sigma}_e^2$ of the residuals, the estimated variances $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_q^2$ of the random effects and the mean squared error mse calculated as an ℓ^2 error rate between the reality -$X\beta$- and the estimation -$X\hat{\beta}$-. We also calculated the Signal-to-Noise Ratio (SNR) as $||X\beta||_2^2 / || \sum_{k=1}^q Z_k u_k + \epsilon||_2^2$ for each of the replications.

4.3 Comments on the results

The detailed results of the simulation study are available in Web Appendix A. A summary of the main results is shown in Figure 1 ($\alpha = 0.1$ for the *procbol* method and the *pbol+* method). No results are given for the *lmmLasso* of Schelldorfer et al. (2011) in Model M_3 since two different grouping factors are considered and the R-package lmmLasso does not include that setting.

In all models, there is an improvement of the results when we switch from a simple linear model to a linear mixed model; indeed there is a significant difference between *Lasso* and *Lasso+* or *procbol* and *pbol+*, especially with model M_4.

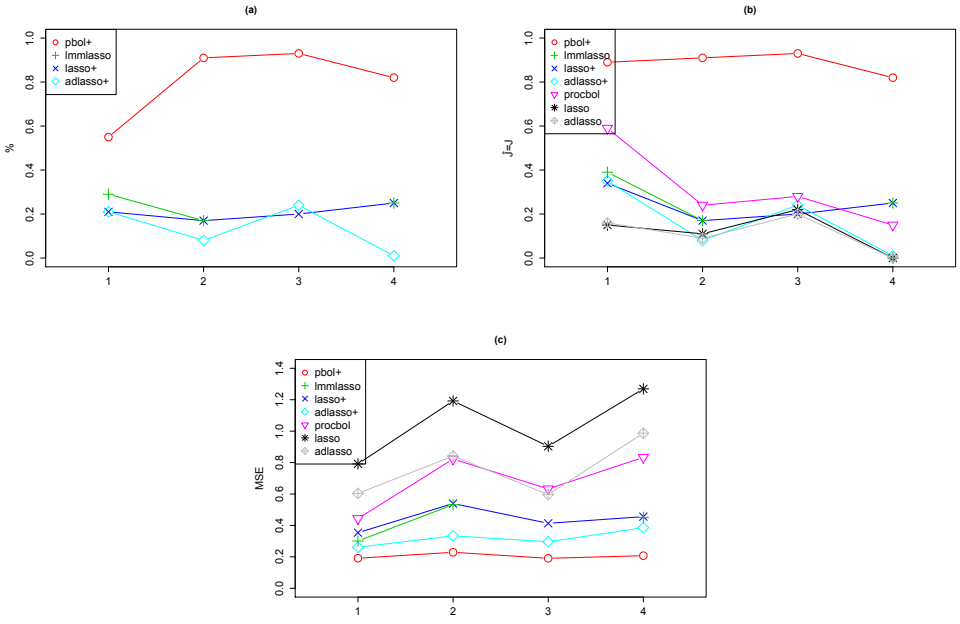


Figure 1: Summary of the results of the simulation study for models $M_1 - M_4$ (X axis). Results of 'Truth' (a), '$\hat{J} = J$' (b) and Mean Squared Error (c) for each model.

On all models, *lmmLasso* and *Lasso+* give very similar results; this is not surprising since both are a ℓ^1-penalization of the log likelihood, except for model M_1 where *lmmLasso* seems to give better results. This difference comes from the coding of the R-package that contains the *lmmLasso* method. Indeed, a variable that generates both a fixed and a random effect does not undergo feature selection in the *lmmLasso* method when the random effect tends towards zero, whereas the *Lasso+* method would allow it.
We observed on our simulation study that both *lmmLasso* and *Lasso+* are very sensitive to the choice of the regularization parameter. On most simulations of model M_4 in which $p = 600$, we observed an edge effect between a regularization parameter that selects few fixed effects (fewer than 15) and a regularization parameter that selects too much fixed-effects ($|\hat{J}| > n$) and thus stops the algorithm because we assumed that the number of relevant fixed-effects is lower than $\min(n-1, p)$, see Figure 4.3. Nevertheless, the weights included in the *adLasso+* seems to smooth this phenomenon, see Figure 4.3 for the same simulation as Figure 4.3. Remark that for the run of model M_4 which is on Figure 2, *Lasso+* could select the true model for a regularization parameter around 0.22 whereas *adLasso+* could not as a noisy variable enters the set of selected variables before all the relevant fixed-effects do.

Concerning the *adLasso+* method, it appears to improve the *Lasso+* method, except

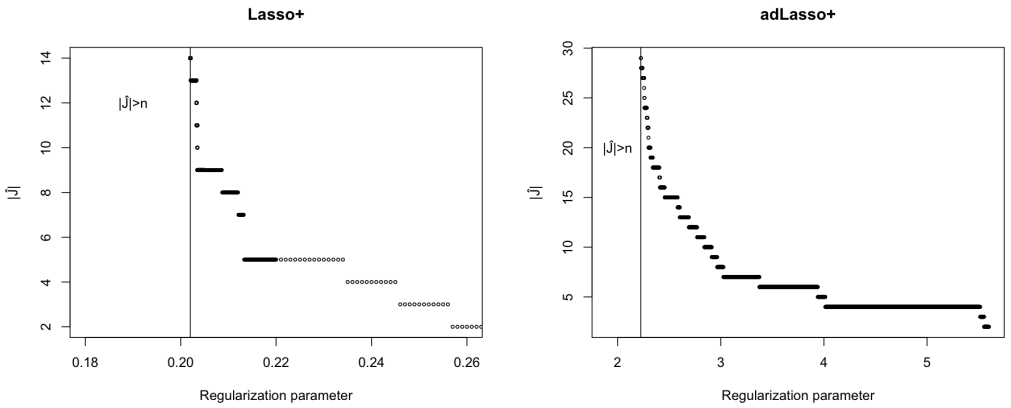


Figure 2: Number of selected fixed effects coefficients depending on the value of the regularization parameter for one run of model M4, for the method (a) *Lasso+* and (b) *adLasso+*. The grid of the penalty is as thin as 10^{-7} next to the area $|\hat{J}| > n$ in (a) and 10^{-3} in (b).

for model M_4 where the true model is only selected once over the 100 replications. On this particular model M_4, *adLasso+* selects more fixed effects but less relevant ones than *Lasso+*. This could mean that the initial weights are not adapted to this case. Despite the result of 'Truth', the *mse* is lower for *adLasso+* than for *Lasso+*.

Algorithm 3.1 combined with the procbol method (*pbol+*) gives the best results over all tested methods for all models. Indeed the percentage of true model recovered is the largest over all methods, the estimation of the fixed effects is really close to the reality and the *mse* is the lowest among the tested methods. Nevertheless, due to the bias of the Lasso, the results in term of *mse* for *Lasso+* and *lmmLasso* could easily be improved with a linear mixed model estimation as said in Section 2.3 (see Web Appendix). Yet, the results of *pbol+* are mitigated for model M_1. Indeed, the percentage of true model recovered is lower than in the other models because of the selection of the random effects that lacks efficiency (the results concerning the selection of the fixed-effects are equivalent as in the other models, as shown in Figure 1). Nonetheless, the results are still better than for the others methods. Moreover, a relevant random effect was never falsely deleted in all models and for all methods. It is interesting to note that the *pbol+* method always converged on our simulations.

A R-package "MMS" is available on CRAN (http://cran.r-project.org). This package contains tools to perform fixed effects selection in linear mixed models; it contains the previous methods denoted as *Lasso+*, *adLasso+*, *pbol+*, among others.
All the results presented in this section were obtained with a specific initialization of the algorithms. The next paragraph is dedicated to the analysis of the influence of that specific initialization.

4.4 Influence of the initialization of our algorithms

Both Algorithm 2.1 and Algorithm 3.1 start with an initialization of the parameter $\Phi = (\sigma_1^2, \ldots, \sigma_q^2, \sigma_e^2, \beta)$. We choose to initialize each algorithm with the following setting: for all $1 \leq k \leq q, \sigma_k^{2[0]} = \frac{0.4}{q}\ \sigma_e^{2[-1]}, \sigma_e^{2[0]} = 0.6\ \sigma_e^{2[-1]}$, and $(\sigma_e^{2[-1]}, \beta^{[0]})$ is estimated from a linear estimation (without the random effects) of the method used at Step 2.

In the current Section, we choose different initializations of Algorithm 2.1 and Algorithm 3.1, both on Model M_4 (see Section 4). The initial values of the variances were set from 0.1 to 10 and of the fixed effects coefficients from -100 to 100. Each algorithm always converged towards the same point, whatever the initialization of Φ, not shown. However, the farther $\Phi^{[0]}$ is set from the true estimation of Φ, the higher is the number of iterations of the algorithms.

5 Application on a real data-set

In this section we analyze a real data set which comes from Rohart et al. (2012). The aim of this analysis is to pinpoint metabolomic data that describes a phenotype taking into account all the available information such as the breed, the batch effect and the relationship between individuals. Here we will study the Daily Feed Intake phenotype (DFI). We model the data as follows:

$$y = X_B\beta_B + X_M\beta_M + Z_E u_E + Z_F u_F + \epsilon, \tag{7}$$

where y is the DFI phenotype, X_B, X_M, Z_E, Z_F are the design matrices of the breed effect, the metabolomic data, the batch effect and the family effect, respectively. We consider two random effects: the batch and the family, considering that each level of these factors is a random sample drawn from a much larger population of batches and families, contrary to the breed factor. Note that the coefficients β_B do not undergo feature selection.
We compare several methods on this model: *Lasso*, *adLasso*, *procbol*, *Lasso+*, *adLasso+* and *pbol+* (see Section 4). The model which is considered for the first three methods is $y = X_B\beta_B + X_M\beta_M + \epsilon$. Both methods *procbol* and *pbol+* were set with a user-level of $\alpha = 0.1$. The results are presented in Table 1.
We observe that considering random effects leads to a decrease of both the residual variance and the number of selected metabolomic variables. This behavior is in accordance with the simulation study. The question that arises from this analysis is to know whether the variables which are selected in the linear mixed models are more relevant than in the linear model. Biological analyses remain to be done to answer that question.

Table 2 gives the computational time of one run when we only consider the batch effect -in order to be able to compute the *lmmLasso*-, showing that the *Lasso+* method is much faster than the *lmmLasso* method for a large number of observations (due to the inversion

	$\lvert\hat{J}\rvert$	$\hat{\sigma}_e^2$	$\hat{\sigma}_E^2$	$\hat{\sigma}_F^2$
Lasso	14	3.8×10^{-2}	-	-
adLasso	21	3.4×10^{-2}	-	-
procbol	11	4.1×10^{-2}	-	-
Lasso+	11	3.2×10^{-2}	3.2×10^{-3}	6.4×10^{-3}
adLasso+	10	3.3×10^{-2}	2.5×10^{-3}	6.5×10^{-3}
pbol+	5	3.4×10^{-2}	5.9×10^{-3}	6.5×10^{-3}

Table 1: Results for the real data set

Methods	CPU Time
Lasso+	0.80
lmmLasso	24.28

Table 2: CPU Time on a single run that selects the same model

of the matrix of variance V at each step of the convergence process). The simulation was performed at a regularization parameter that selects the same model for the two methods, on a 2.80GHz CPU with 8.00Go of RAM.

6 Conclusion

In this paper, we proposed to add a ℓ^1-penalization of the complete log-likelihood in order to perform selection of the fixed effects in a linear mixed model. The multicycle ECM algorithm used to minimize the objective function also performs random effects selection. This algorithm gives the same results as the lmmLasso of Schelldorfer et al. (2011) when the random effects are assumed to be independent, but faster. Theoretical results are identical to those of Schelldorfer et al. (2011) when the variances are known. The structure of our algorithm gives the possibility to combine it with any variable selection method built for linear models, but at the price of possibly loosing the convergence property. Nonetheless, the combined procbol method appears to give good results on simulated data and outperforms other approaches.
We applied all these methods to a real data set showing that the residual variance can be reduced, even with a small set of selected variables.

Supplementary Materials

Web Appendices, referenced in Section 2 and 4, are available with this paper at the Biometrics website on Wiley Online Library.

References

Bach, F. (2009). Model-consistent sparse estimation through the bootstrap. Technical report, hal-00354771, version 1.

Biernacki, C. and Chrétien, S. (2003). Degeneracy in the maximum likelihood estimation of univariate gaussian mixtures with em. *Statistics & Probability Letters*, 61:373–382.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection of fixed and random effects in linear mixed-effects models. *Biometrics*, 66:1069–1077.

Foulley, J. (1997). Ecm approaches to heteroskedastic mixed models with constant variance ratios. *Genetics Selection Evolution*, 29:197–318.

Foulley, J.-L., Delmas, C., and Robert-Granié, C. (2006). Méthodes du maximum de vraisemblance en modèle linéaire mixte. *J. SFdS*, 1-2:5–52.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, 72:320–340.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9:226–252.

Henderson, C. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, pages 10–41.

Henderson, C. (1984). *Applications of linear models in Animal breeding*. University of Guelph, Ont.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptative lasso for sparse high-dimensional regression models. *Stat. Sin.*, 18(4):1603–1618.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67:495–503.

McLachlan, J. and Krishnan, T. (2008). *The EM Algorithm and Extensions, second edition*. Wiley-Interscience.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80:267–278.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Rohart, F. (2011). Multiple hypotheses testing for variable selection. *arXiv:1106.3415v1*.

Rohart, F., Paris, A., Laurent, B., Canlet, C., Molina, J., Mercat, M. J., Tribout, T., Muller, N., Ianuccelli, N., Villa-Vialaneix, N., Liaubet, L., Milan, D., and San-Cristobal, M. (2012). Phenotypic prediction based on metabolomic data on the growing pig from three main european breeds. *Journal of Animal Science*.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scand. J. Stat.*, 38:197–214.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist*, 6(2):461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.*, B 58(1):267–288.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc.*, B 68:46–67.

Zhang, C.-H. and Hunag, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J.R. Statist. Soc.*, B 67(2):301–320.

Web Appendix for "Fixed effects Selection in high dimensional Linear Mixed Models"

Florian Rohart[1,2], Magali San-Cristobal [2] and Béatrice Laurent [1]
[1] UMR 5219, Institut de Mathématiques de Toulouse,
INSA de Toulouse, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France

[2] UMR 444 Laboratoire de Génétique Cellulaire,
INRA Toulouse, 31320 Castanet Tolosan cedex, France

2012

Web Appendix A - Results of the simulation study

Table 1: Results of model M_1. The percentage of true model recovered was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 0.78(0.13)$. Standard errors are given in parentheses, for 100 runs.

	Truth	$\hat{J} = J$	$\lvert\hat{J}\rvert$	TP	$\hat{\sigma}_e^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\sigma}_3^3$
Ideal	1	5	5	5	1	1	1	0
Lasso	-	0.15	4.95	4.13	3.27	-	-	-
			(1.90)	(1.12)	(0.62)	-	-	-
adLasso	-	0.16	5.25	4.26	2.91	-	-	-
			(1.84)	(0.89)	(0.59)	-	-	-
procbol	-	0.59	4.70	4.58	2.83	-	-	-
$\alpha = 0.1$			(0.78)	(0.61)	(0.57)	-	-	-
procbol	-	0.45	4.47	4.40	2.89	-	-	-
$\alpha = 0.05$			(0.67)	(0.62)	(0.58)	-	-	-
Lasso+	0.21	0.34	6.42	5.00	1.04	0.88	0.98	0.02
			(1.64)	(0.00)	(0.21)	(0.37)	(0.44)	(0.06)
adLasso+	0.21	0.35	6.34	4.99	0.94	0.86	0.95	0.02
			(1.41)	(0.10)	(0.18)	(0.36)	(0.41)	(0.06)
lmmLasso	0.29	0.39	6.15	5.00	1.01	0.89	0.96	0.02
			(1.29)	(0.00)	(0.19)	(0.38)	(0.42)	(0.06)
pbol+	0.55	0.89	5.18	5.00	0.92	0.87	0.97	0.03
$\alpha = 0.1$			(0.50)	(0.00)	(0.18)	(0.37)	(0.41)	(0.06)
pbol+	0.59	0.93	5.08	5.00	0.93	0.88	0.97	0.03
$\alpha = 0.05$			(0.30)	(0.00)	(0.17)	(0.37)	(0.41)	(0.06)

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	MSE
Ideal	0.67	0.67	0.67	0.67	0.67	0.00
Lasso	0.67	0.29	0.31	0.41	0.17	0.79
	(0.27)	(0.26)	(0.20)	(0.19)	(0.16)	(0.42)
adLasso	0.69	0.42	0.46	0.58	0.27	0.60
	(0.27)	(0.33)	(0.25)	(0.23)	(0.22)	(0.37)
procbol	0.69	0.63	0.68	0.65	0.49	0.44
$\alpha = 0.1$	(0.27)	(0.32)	(0.17)	(0.30)	(0.33)	(0.31)
procbol	0.69	0.63	0.68	0.62	0.43	0.51
$\alpha = 0.05$	(0.27)	(0.32)	(0.17)	(0.33)	(0.36)	(0.30)
Lasso+	0.69	0.65	0.49	0.41	0.43	0.35
	(0.25)	(0.28)	(0.17)	(0.11)	(0.11)	(0.17)
adLasso+	0.69	0.64	0.59	0.57	0.48	0.26
	(0.25)	(0.27)	(0.15)	(0.12)	(0.14)	(0.15)
lmmLasso	0.69	0.65	0.66	0.41	0.43	0.30
	(0.25)	(0.28)	(0.11)	(0.11)	(0.10)	(0.15)
pbol+	0.69	0.67	0.67	0.66	0.65	0.19
$\alpha = 0.1$	(0.25)	(0.28)	(0.12)	(0.10)	(0.10)	(0.14)
pbol+	0.69	0.67	0.67	0.66	0.65	0.18
$\alpha = 0.05$	(0.25)	(0.28)	(0.11)	(0.10)	(0.10)	(0.13)

Table 2: Results of model M_2. The percentage of true model recovered was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 1.26(0.25)$. Standard errors are given in parentheses, for 100 runs.

Results	Truth	$\hat{J} = J$	$\lvert\hat{J}\rvert$	TP	$\hat{\sigma}_e^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
Ideal	1	5	5	5	1	1	1
Lasso	-	0.11	5.02	3.86	3.62	-	-
			(2.69)	(1.35)	(0.96)	-	-
adLasso	-	0.09	6.06	4.24	3.05	-	-
			(2.66)	(1.16)	(0.87)	-	-
procbol	-	0.24	3.95	3.76	3.62	-	-
$\alpha = 0.1$			(1.22)	(1.06)	(0.95)	-	-
procbol	-	0.21	3.60	3.47	3.53	-	-
$\alpha = 0.05$			(1.25)	(1.14)	(0.87)	-	-
Lasso+	0.17	0.17	7.60	4.92	1.25	0.91	0.93
			(2.64)	(0.37)	(0.28)	(0.40)	(0.48)
adLasso+	0.08	0.08	8.26	5.00	0.99	0.90	0.85
			(3.15)	(0.00)	(0.21)	(0.38)	(0.41)
lmmLasso	0.17	0.17	7.65	4.93	1.24	0.91	(0.93)
			(2.49)	(0.36)	(0.26)	(0.40)	(0.48)
pbol+	0.91	0.91	4.86	4.85	1.01	0.95	0.88
$\alpha = 0.1$			(0.59)	(0.58)	(0.28)	(0.38)	(0.41)
pbol+	0.80	0.80	4.57	4.57	1.11	0.93	0.88
$\alpha = 0.05$			(0.93)	(0.93)	(0.39)	(0.38)	(0.39)

	$\hat{\beta}_{i_1}$	$\hat{\beta}_{i_2}$	$\hat{\beta}_{i_3}$	$\hat{\beta}_{i_4}$	$\hat{\beta}_{i_5}$	MSE
Ideal	0.75	0.75	0.75	0.75	0.75	0.00
Lasso	0.79	0.47	0.21	0.19	0.17	1.19
	(0.27)	(0.31)	(0.19)	(0.17)	(0.16)	(0.57)
adLasso	0.79	0.64	0.36	0.35	0.29	0.84
	(0.27)	(0.38)	(0.24)	(0.24)	(0.22)	(0.55)
procbol	0.79	0.72	0.50	0.57	0.52	0.82
$\alpha = 0.1$	(0.27)	(0.49)	(0.40)	(0.38)	(0.38)	(0.55)
procbol	0.79	0.75	0.44	0.50	0.45	0.93
$\alpha = 0.05$	(0.27)	(0.50)	(0.42)	(0.41)	(0.40)	(0.56)
Lasso+	0.82	0.91	0.35	0.35	0.33	0.54
	(0.26)	(0.26)	(0.13)	(0.11)	(0.13)	(0.24)
adLasso+	0.81	0.82	0.51	0.52	0.49	0.33
	(0.25)	(0.25)	(0.14)	(0.13)	(0.14)	(0.17)
lmmLasso	0.82	0.91	0.35	0.35	0.33	0.53
	(0.26)	(0.26)	(0.13)	(0.11)	(0.13)	(0.23)
pbol+	0.79	0.76	0.70	0.73	0.72	0.23
$\alpha = 0.1$	(0.25)	(0.26)	(0.22)	(0.17)	(0.18)	(0.28)
pbol+	0.80	0.79	0.64	0.66	0.66	0.35
$\alpha = 0.05$	(0.25)	(0.28)	(0.29)	(0.28)	(0.28)	(0.43)

Table 3: Results of model M_3. The percentage of true model recovered was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 0.83(0.16)$. Standard errors are given in parentheses, for 100 runs.

Results	Truth	$\hat{J} = J$	$\lvert\hat{J}\rvert$	TP	$\hat{\sigma}_e^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
Ideal	1	5	5	5	1	1	1
Lasso	-	0.22	4.96	4.13	3.32	-	-
			(2.18)	(1.10)	(0.80)	-	-
adLasso	-	0.20	6.10	4.58	2.85	-	-
			(2.19)	(0.70)	(0.72)	-	-
procbol	-	0.28	4.37	4.12	2.90	-	-
$\alpha = 0.1$			(1.08)	(0.77)	(0.79)	-	-
procbol	-	0.26	4.17	3.97	2.97	-	-
$\alpha = 0.05$			(1.12)	(0.83)	(0.82)	-	-
Lasso+	0.20	0.20	7.07	4.99	1.11	0.91	0.92
			(2.01)	(0.10)	(0.22)	(0.36)	(0.46)
adLasso+	0.24	0.24	6.70	4.97	0.97	0.88	0.88
			(1.51)	(0.17)	(0.19)	(0.34)	(0.45)
lmmLasso	-	-	-	-	-	-	-
			-	-	-	-	-
pbol+	0.93	0.93	5.09	5.00	0.95	0.91	0.89
$\alpha = 0.1$			(0.38)	(0.00)	(0.17)	(0.33)	(0.44)
pbol+	0.95	0.95	5.08	5.00	0.95	0.91	0.89
$\alpha = 0.05$			(0.44)	(0.00)	(0.17)	(0.33)	(0.44)

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	MSE
Ideal	0.67	0.67	0.67	0.67	0.67	0.00
Lasso	0.69	0.69	0.18	0.20	0.27	0.90
	(0.25)	(0.32)	(0.17)	(0.17)	(0.17)	(0.40)
adLasso	0.69	0.68	0.32	0.36	0.46	0.60
	(0.25)	(0.32)	(0.21)	(0.21)	(0.22)	(0.32)
procbol	0.73	0.65	0.48	0.51	0.57	0.63
$\alpha = 0.1$	(0.34)	(0.13)	(0.36)	(0.36)	(0.35)	(0.42)
procbol	0.73	0.65	0.44	0.49	0.56	0.68
$\alpha = 0.05$	(0.34)	(0.13)	(0.38)	(0.38)	(0.36)	(0.43)
Lasso+	0.71	0.71	0.40	0.38	0.43	0.41
	(0.24)	(0.29)	(0.12)	(0.11)	(0.11)	(0.19)
adLasso+	0.71	0.69	0.50	0.48	0.56	0.30
	(0.24)	(0.29)	(0.16)	(0.14)	(0.13)	(0.18)
lmmLasso	-	-	-	-	-	-
	-	-	-	-	-	-
pbol+	0.71	0.69	0.67	0.65	0.68	0.19
$\alpha = 0.1$	(0.24)	(0.29)	(0.12)	(0.10)	(0.10)	(0.16)
pbol+	0.71	0.69	0.67	0.65	0.68	0.19
$\alpha = 0.05$	(0.24)	(0.29)	(0.12)	(0.10)	(0.10)	(0.16)

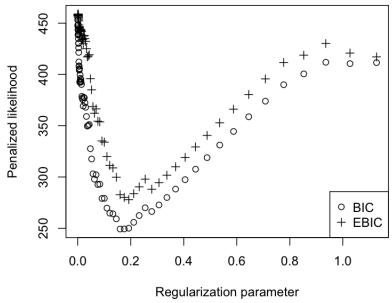
Table 4: Results of model M_4. The percentage of true model recovered was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 0.63(0.11)$. Standard errors are given in parentheses, for 100 runs.

Results	Truth	$\hat{J} = J$	$\lvert\hat{J}\rvert$	TP	$\hat{\sigma}_e^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
Ideal	1	5	5	5	1	1	1
Lasso	-	0.00	2.81	2.06	4.08	-	-
			(2.80)	(1.30)	(0.84)	-	-
adLasso	-	0.00	5.64	3.03	3.38	-	-
			(4.10)	(1.22)	(0.88)	-	-
procbol	-	0.15	3.85	3.61	3.23	-	-
$\alpha = 0.1$			(1.00)	(0.95)	(0.73)	-	-
procbol	-	0.15	3.48	3.34	3.39	-	-
$\alpha = 0.05$			(1.00)	(0.99)	(0.80)	-	-
Lasso+	0.25	0.25	7.13	4.99	1.21	0.93	1.03
			(1.84)	(0.10)	(0.27)	(0.41)	(0.40)
adLasso+	0.01	0.01	9.56	4.87	0.94	0.89	0.98
			(4.01)	(0.37)	(0.26)	(0.37)	(0.37)
lmmLasso	0.25	0.25	7.22	4.99	1.19	0.93	1.03
			(1.95)	(0.10)	(0.25)	(0.40)	(0.40)
pbol+	0.82	0.82	5.21	4.99	0.92	0.97	1.00
$\alpha = 0.1$			(0.56)	(0.10)	(0.17)	(0.39)	(0.34)
pbol+	0.88	0.88	5.10	4.98	0.93	0.97	1.00
$\alpha = 0.05$			(0.41)	(0.14)	(0.16)	(0.39)	(0.34)

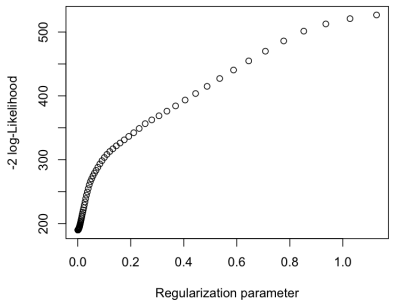
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	MSE
Ideal	0.67	0.67	0.67	0.67	0.67	0.00
Lasso	0.60	0.06	0.06	0.06	0.11	1.27
	(0.25)	(0.15)	(0.11)	(0.11)	(0.15)	(0.32)
adLasso	0.60	0.15	0.18	0.17	0.26	0.99
	(0.25)	(0.26)	(0.19)	(0.19)	(0.23)	(0.33)
procbol	0.60	0.55	0.38	0.44	0.43	0.83
$\alpha = 0.1$	(0.25)	(0.31)	(0.38)	(0.39)	(0.40)	(0.43)
procbol	0.60	0.53	0.32	0.38	0.35	0.91
$\alpha = 0.05$	(0.25)	(0.32)	(0.38)	(0.40)	(0.41)	(0.39)
Lasso+	0.62	0.55	0.31	0.35	0.37	0.46
	(0.25)	(0.27)	(0.11)	(0.12)	(0.12)	(0.20)
adLasso+	0.61	0.56	0.41	0.43	0.48	0.39
	(0.25)	(0.26)	(0.16)	(0.18)	(0.16)	(0.19)
lmmLasso	0.62	0.55	0.31	0.35	0.38	0.45
	(0.25)	(0.27)	(0.11)	(0.12)	(0.12)	(0.19)
pbol+	0.60	0.64	0.67	0.67	0.67	0.21
$\alpha = 0.1$	(0.25)	(0.28)	(0.10)	(0.11)	(0.13)	(0.15)
pbol+	0.60	0.64	0.67	0.67	0.67	0.20
$\alpha = 0.05$	(0.25)	(0.27)	(0.10)	(0.13)	(0.13)	(0.15)

Table 5: Results of model M_4 when a ML linear regression is added after the convergence of the algorithm. The percentage of true model recovered was recorded -'Truth'- as well as $\hat{J} = J$. $|J|$ is the number of fixed effects selected and TP the number of relevant fixed effects selected. The signal to noise ratio is equal to $SNR = 0.63(0.11)$. Standard errors are given in parentheses, for 100 runs.

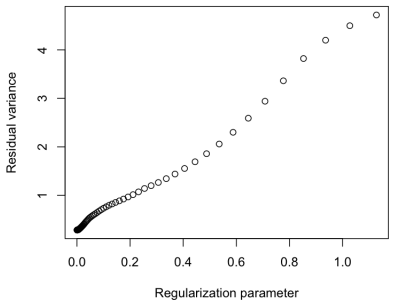
	Ideal	lmmLasso	Lasso+
Truth	1	0.25	0.25
$\hat{J} = J$	1	0.25	0.25
$\lvert\hat{J}\rvert$	5	7.22(1.95)	7.13(1.84)
TP	5	4.99(0.10)	4.99(0.10)
$\hat{\sigma}_e^2$	1	1.19(0.25)	1.21(0.27)
$\hat{\sigma}_1^2$	1	0.96(0.39)	0.96(0.40)
$\hat{\sigma}_2^2$	1	1.01(0.36)	1.01(0.36)
$\hat{\beta}_1$	0.67	0.61(0.25)	0.61(0.25)
$\hat{\beta}_2$	0.67	0.62(0.28)	0.62(0.28)
$\hat{\beta}_3$	0.67	0.61(0.12)	0.61(0.12)
$\hat{\beta}_4$	0.67	0.63(0.12)	0.63(0.12)
$\hat{\beta}_5$	0.67	0.62(0.14)	0.62(0.14)
mse	0	0.40(0.17)	0.40(0.17)



(a) BIC or EBIC depending on the value of the regularization parameter of the Lasso method



(b) $-2\times$log-Likelihood depending on the regularization parameter of the Lasso method



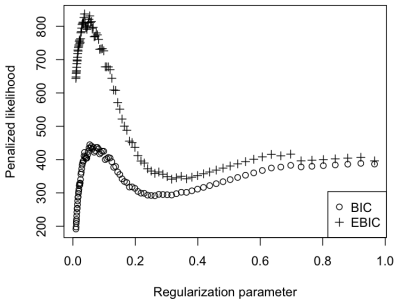
(c) Residual variance depending on the regularization parameter of the Lasso method

Figure 1: One simulation of linear model for the Lasso method with $n = 120, p = 80$ and $\beta_J = 1$.

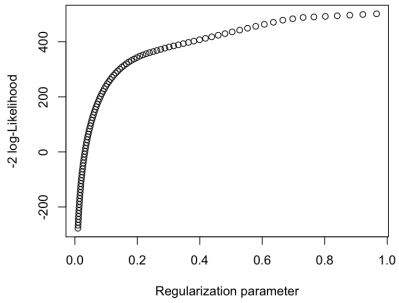
Web Appendix B - Remark on the tuning parameter

The tuning of the regularization parameter could be tricky for some methods, especially the *Lasso* method and the *adLasso* method. In this section, we look at the causes.

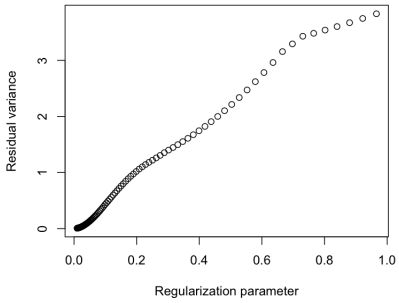
We shall begin to consider the classical linear model before studying the linear mixed model. Let us first look at the Lasso method when only applied in a classical linear model. We compare two penalizations of the likelihood: BIC and the Extended BIC (EBIC) (Chen and Chen, 2008). The EBIC penalizes a space of dimension k with a term that depends on the number of spaces that have the same dimension, which is $\frac{p!}{k!(p-k)!}$; thus EBIC penalizes more the complex spaces than BIC. Figure 1 shows the behavior of the BIC and EBIC criteria, the log-likelihood and the residual variance for several values of the regularization parameter of the Lasso in a low dimensional case ($p = 80$). We observe that tuning the regularization parameter in this case raises no problem.



(a) BIC or EBIC depending on the value of the regularization parameter of the Lasso



(b) $-2\times$log-Likelihood depending on the regularization parameter of the Lasso method



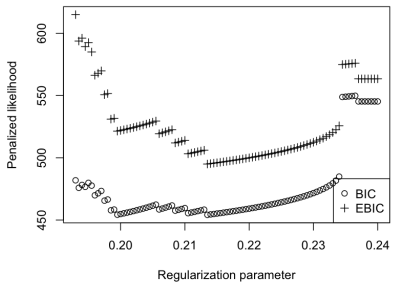
(c) Residual variance depending on the regularization parameter of the Lasso method

Figure 2: One simulation of linear model for the Lasso method with $n = 120, p = 600$ and $\beta_J = 1$.

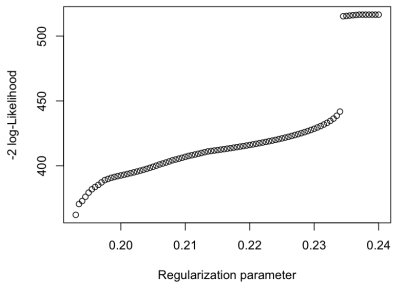
Let us now consider a simulation in a high dimensional context in which we have $n = 120$ observations and $p = 600$ explanatory variables. Results of the two methods for choosing the regularization parameter of Lasso are presented in Figure 2.

Firstly, we confirm that EBIC is more conservative than BIC and penalizes more the complex spaces. On the far left of Figure 2(a), we observe that both the BIC and the EBIC curves decrease when the regularization parameter is close to zero. This phenomenon is due to the degeneracy of the likelihood that can be seen in Figure 2(b) (stated in Section **??** for mixed models, it can also happen in linear models). Figure 2(c) shows that the degeneracy of the likelihood comes from the residual variance that drops to zero when the regularization parameter is close to zero, and thus when too much variables enter the model.

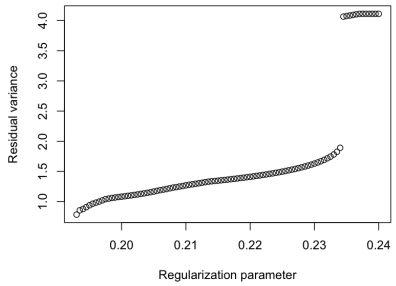
To conclude, we see that both BIC and EBIC penalties are not sufficiently strong to completely balance the degeneracy of the likelihood; however, EBIC penalty leads to select



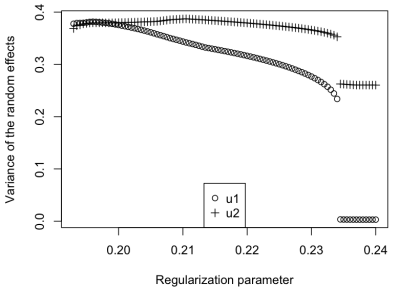
(a) BIC or EBIC depending on the value of the regularization parameter of the Lasso+ method



(b) $-2\times$log-Likelihood depending on the regularization parameter of the Lasso+ method



(c) Residual variance depending on the regularization parameter of the Lasso+ method



(d) Residual variance depending on the regularization parameter of the Lasso+ method

Figure 3: One simulation of linear mixed model with $n = 120, p = 600, \beta_J = 1$ and two i.i.d. random effects.

a more parsimonious model while BIC penalty selects a more complex model. Nonetheless, the EBIC penalty is usually too much conservative in practice, that is why the simulation study used the BIC penalty. When the degeneracy happens -as it is likely to occur as p grows-, the regularization parameter should be optimized over an area that does not contain the explosion of the likelihood, that means that the area should not contain the far left part of Figure 2(a) where the criterion decreases.

We now look at the *Lasso+* method. As mentioned in the paper, the maximal number of fixed-effects that can be selected with the *Lasso+* method is small compared to n or p. Thus, the degeneracy of the likelihood never occurred in our simulations (Figure 3). However, if this phenomenon happens, the choice of the grid of the regularization parameter should follow the same advice as the one given above for the classical linear model.

Web Appendix C - Proof of Proposition 2.2

G and R are supposed to be known. Thus the minimization of our objective function g reduces to the minimization of the following function in (β, u):
$h(u,\beta) = (y - X\beta - Zu)'R^{-1}(y - X\beta - Zu) + u'G^{-1}u + \lambda|\beta|_1$.
Let denote $(\hat{u}, \hat{\beta}) = \underset{(u,\beta)}{\text{argmin}}\ h(u,\beta)$. Since the function h is convex, we have:

$$(\hat{u}, \hat{\beta}) = \begin{cases} u(\beta) = \underset{u}{\text{argmin}}\ h(u,\beta) \\ \hat{\beta} = \underset{\beta}{\text{argmin}}\ h(u(\beta),\beta) \\ \hat{u} = u(\hat{\beta}) \end{cases}.$$

Since $\dfrac{\partial h(u,\beta)}{\partial u}$ exists, we can explicit the minimum of h in u:

$$(\hat{u}, \hat{\beta}) = \begin{cases} u(\beta) = (Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - X\beta) \\ \hat{\beta} = \underset{\beta}{\text{argmin}}\ h(u(\beta),\beta) \\ \hat{u} = u(\hat{\beta}) \end{cases}$$

Thus, we obtain:

$$\begin{aligned} h(u(\beta),\beta) &= (y - X\beta - Zu(\beta))'R^{-1}(y - X\beta - Zu(\beta)) + u'G^{-1}u + \lambda|\beta|_1 \\ &= (y - X\beta)'R^{-1}(y - X\beta) - (y - X\beta)R^{-1}Zu(\beta) - (Zu(\beta))'R^{-1}(y - X\beta) \\ &\quad + (Z\hat{u})'R^{-1}Zu(\beta) + u(\beta)'G^{-1}u(\beta) + \lambda|\beta|_1 \\ &= (y - X\beta)'\left[R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}\right](y - X\beta) + \lambda|\beta|_1 \end{aligned}$$

Denote $W = R^{-1} - R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}$. We can show that $W = (Z'GZ + R^{-1})^{-1} = V^{-1}$. This result comes from the equivalence between the resolution of Henderson's equations (Henderson, 1973) and the generalized least squares.

To conclude, we have that

$$(\hat{u}, \hat{\beta}) = \left((Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - X\hat{\beta}), \underset{\beta}{\text{argmin}}\ (y - X\beta)'V^{-1}(y - X\beta) + \lambda|\beta|_1\right).$$

References

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 94:759–771.

Henderson, C. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, pages 10–41.

4.3 Conclusions

Une nouvelle méthode de sélection d'effets fixes dans un modèle linéaire mixte a été présentée dans la partie précédente. Cette méthode donne des résultats très satisfaisants sur les simulations aussi bien en petite dimension qu'en grande dimension. L'algorithme utilisé pour cette méthode se combine aisément aux méthodes de sélection de variables existantes dans le modèle linéaire classique. La combinaison de cet algorithme avec la procédure de tests multiples présentée dans la Partie 3, *procbol+*, donne de bons résultats en simulation ainsi que sur les données réelles. Des résultats théoriques sur la consistance de notre méthode *Lasso+* dans le cas particulier où les variances sont connues ont été donnés ; un travail complémentaire est à accomplir afin d'obtenir des résultats théoriques pour le cas général.

Comparons la liste des métabolites sélectionnés pour le modèle linéaire mixte et le modèle linéaire classique pour le phénotype DFI dans le modèle qui prend en compte la race des individus, les résultats sont donnés en Table 7.

Modèle linéaire				Modèle linéaire mixte			
Lasso		procbol		Lasso+		procbol+	
δ (n)	Assign.	δ (n)	Assign.	δ(n)	Assign.	δ (n)	Assign.
4.05 (100)	creatinine	4.05 (96)	creatinine	4.05 (97)	creatinine	4.05 (100)	creatinine
2.04 (100)	glutamine			2.04 (97)	glutamine		
	glutamate				glutamate		
	proline				proline		
4.23 (65)	inconnu			0.90 (66)	Lipides		
2.42 (82)	glutamine			2.39 (53)	inconnu		
2.39 (51)	inconnu						
2.26 (76)	valine						
1.90 (80)	inconnu						
1.47 (87)	alanine						
1.46 (60)	alanine						
0.90 (72)	Lipides						
0.84 (81)	Lipides						

TABLE 7 – **Variables sélectionnées pour le phénotype "DFI" pour différentes méthodes. Le décalage chimique (δ) en ppm est donné. Le nombre de fois où la variable est sélectionnée sur les 100 itérations est donné entre parenthèses, seuillé à 50.**

On remarque dans la Table 7 que les métabolites sélectionnés dans le modèle linéaire mixte -que ce soit avec la méthode Lasso+ ou la procédure procbol+- sont aussi sélectionnés dans le modèle linéaire classique -que ce soit avec la méthode Lasso ou la procédure de

tests multiples procbol, respectivement-. D'après les résultats obtenus dans la Section 4.2 pour le phénotype DFI, on observe que le modèle et la méthode qui permettent d'obtenir la plus basse erreur de prédiction sont le modèle linéaire classique et la méthode adLasso. La prise en compte du lien de parenté entre individus ainsi que la prise en compte de la bande en tant qu'effets aléatoires augmente légèrement l'erreur de prédiction pour toutes les méthodes considérées. Ce phénomène peut avoir plusieurs raisons. Il pourrait être dû au plan d'expérience très déséquilibré (entre l'effet race et l'effet bande) ainsi qu'au faible nombre d'individus par famille. Ce dernier point mérite de plus amples investigations. Des travaux complémentaires devraient être menés sur la modélisation de ce jeu de données réelles et sur l'amélioration du plan d'expérience afin de mieux estimer les effets aléatoires.

Cependant, l'objectif de la procédure proposée est d'effectuer une sélection de variables dans un modèle linéaire mixte. Bien entendu sur les données réelles il est impossible de vérifier la performance de la méthode en termes de sélection de variables. Néanmoins, les simulations ont montré que la prise en compte des effets aléatoires améliore de façon substantielle la sélection de variables.

5 Travaux en cours et perspectives

Les perspectives de travail concernent principalement la mise en relation de différents types de données, par exemple des données transcriptomiques et des données génomiques des individus du projet DéLiSus. Les données transcriptomiques ayant été disponibles dans le courant de ce travail, cette thèse a aussi permis l'encadrement d'un projet puis d'un stage de Master 1 sur le sujet "Analyse de données transcriptomiques". Ce travail a porté sur une analyse différentielle afin d'identifier des gènes dont l'expression varie selon les races. Pour se faire, chacune des $p = 12\ 358$ variables transcriptomiques a été analysée dans un modèle linéaire mixte dans lequel la race a été considérée comme effet fixe et la bande des individus en effet aléatoire. Une p-valeur a été calculée dans chacun des p modèles à l'aide du test d'égalité des moyennes de l'effet de chaque race, les p-valeurs ont ensuite été corrigées en contrôlant le taux de faux positifs (FDR) par la méthode de Benjamini and Yekutieli (2001). La Table 8 donne le nombre de transcrits différentiés selon la race pour cette méthode, en fonction du FDR.

FDR	0.1	0.05	0.01	0.001	0.0001
# transcrits	2644	2257	1545	982	610

TABLE 8 – **Nombre de transcrits différentiés entre races pour un ensemble de taux de faux positifs (FDR) fixés.**

Une représentation des individus sur les deux premiers axes ainsi que sur les axes 2 et 3 d'une analyse en composantes principales obtenue à partir des 982 transcrits considérés comme différentiels pour la race avec un seuil FDR de 0.001 est donnée en Figure 8. On observe que l'axe 1 permet de différencier la race Piétrain des autres races. Cependant il est à noter qu'un effet sexe est confondu avec la race Piétrain puisque, dans ce projet et pour des raisons indépendantes de la volonté des scientifiques, les Piétrains sont tous des femelles contrairement aux animaux des autres races qui sont des mâles. L'axe 3 sépare la race Duroc des autres ; les races Landrace et Large White (femelle ou mâle) sont légèrement différenciées suivant l'axe 2.

L'objectif était ici de déterminer un ensemble de gènes différentiellement exprimés pour la race. Toutefois, le problème peut être envisagé comme une question de classification, où l'objectif est de déterminer une liste restreinte de transcrits déterminants dans la différentiation des races et qui permettent de classifier au mieux celles-ci. La méthode Lasso du package *glmnet* pour le logiciel R permet d'effectuer une sélection de variables dans un objectif de classification lorsque la famille des observations est considérée comme multinomiale et que la relation entre les transcrits et la race est supposée linéaire. Notre méthode *Lasso+* et l'algorithme présenté, cf. Section 4, devrait également pouvoir s'adapter aux cas de classification, si le modèle est généralisé. Notons que l'utilisation de méthodes plus classiques comme les forêts aléatoires (Breiman, 2001) ou la sPLS-DA (Lê Cao et al.,

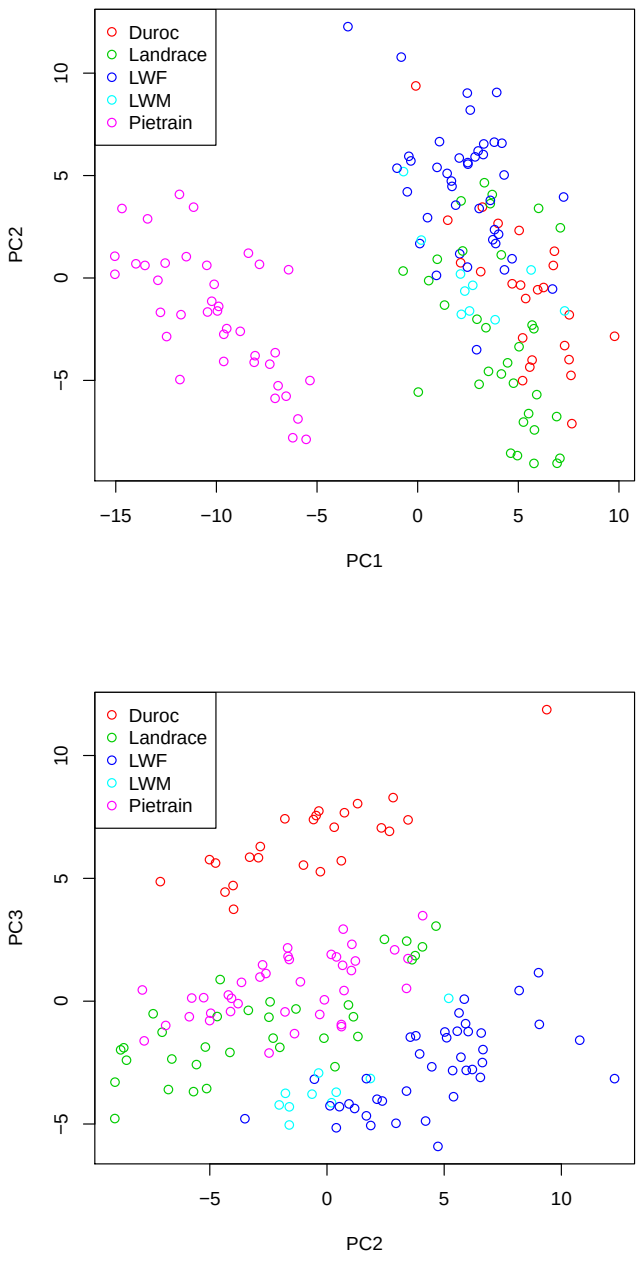


FIGURE 8 – **Représentation des individus sur les premières composantes principales construites avec les 982 transcrits différentiés entre races (FDR<0.001)**

2011) peut aussi être une solution à ce problème.

Par la suite il est intéressant de combiner l'information contenue dans le transcriptome et celle contenue dans le métabolome afin de prédire des phénotypes d'intérêt. Comme mentionné en introduction de ce manuscrit, le coût des expériences est tel que toutes les données n'ont pas été recueillies sur tous les animaux. On obtient donc un nombre d'individus communs dans les deux jeux de données relativement faible comparé au nombre d'observations du métabolome : seulement $n = 102$ individus, décomposés en 39 Large White type femelle, 2 Large White type mâle, 32 Landrace, 23 Piétrain et 6 Duroc. Cette analyse a été conduite sur les 94 individus des trois races majoritaires dans un modèle linéaire par la méthode Lasso. Les résultats sont comparés à la méthode Lasso effectuée sur les 94 mêmes individus lorsque seules les données métabolomiques sont considérées, cf. Figure 9,10 et 11. Le premier constat est une augmentation de la variabilité des erreurs de prédictions sur les trois modèles considérés pour les deux types de données. Cette variabilité peut s'expliquer par la diminution du nombre d'observations et l'accroissement du nombre de paramètres : pour l'analyse couplée du métabolome est du transcriptome on a $p = 12\ 734$ dans le modèle 1 (cf. (12a)), $p = 50\ 936$ dans le modèle 2 (on rajoute la race, cf. (12b)) et $p = 140\ 074$ dans le modèle 3 (on rajoute la race et la bande, cf. (12c)), à mettre en balance avec les $n = 94$ observations. On observe également une amélioration de la prédiction d'une grande majorité de phénotypes dans le modèle le plus simple lorsque le transcriptome est combiné au métabolome, Figure 9. Au contraire, l'ajout du transcriptome n'apporte plus d'information et a tendance à détériorer le pouvoir prédictif du métabolome dans les autres modèles (lorsque la race et la bande sont considérées).
Cette analyse doit être approfondie, notamment par l'utilisation d'autres méthodes de sélection de variables dans le modèle linéaire ainsi que l'utilisation de méthodes de sélection d'effets fixes pour les modèles linéaires mixtes si l'on considère la bande et la relation de parenté comme des effets aléatoires.

En collaboration avec B. Servin (INRA Toulouse), une recherche de mQTL (metabolomic Quantitative Trait Locus) a été entreprise à la fin de cette thèse. Ce type d'analyse permet de mettre en relation les données métabolomiques et les données génomiques - le génome porcin est constitué de 18 chromosomes, les données comportent 46 425 SNP (Single Nucleotide Polymorphism)-.

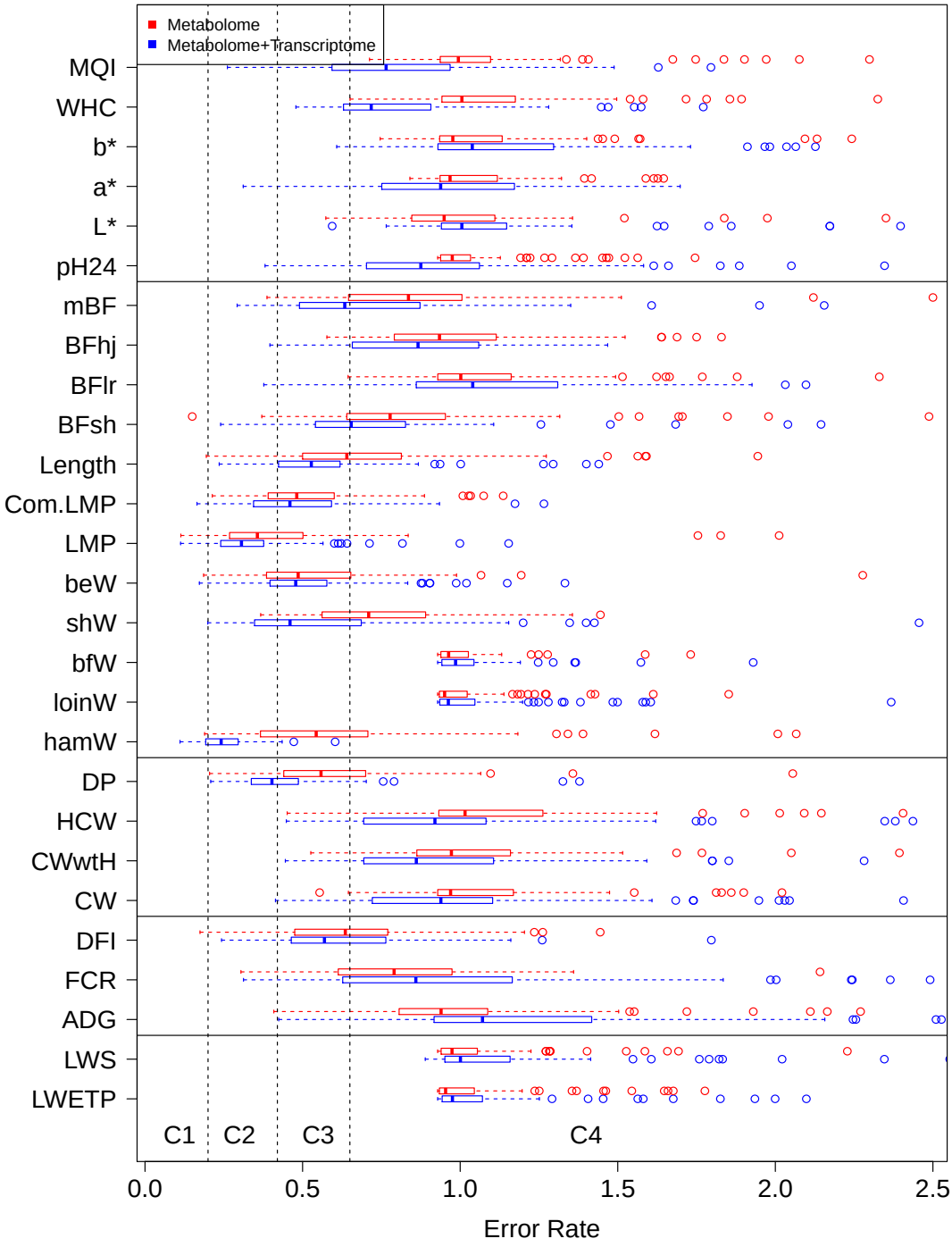


FIGURE 9 – **Prédictions de 27 phénotypes par le métabolome et par la combinaison du métabolome et du transcriptome, avec la méthode Lasso.**

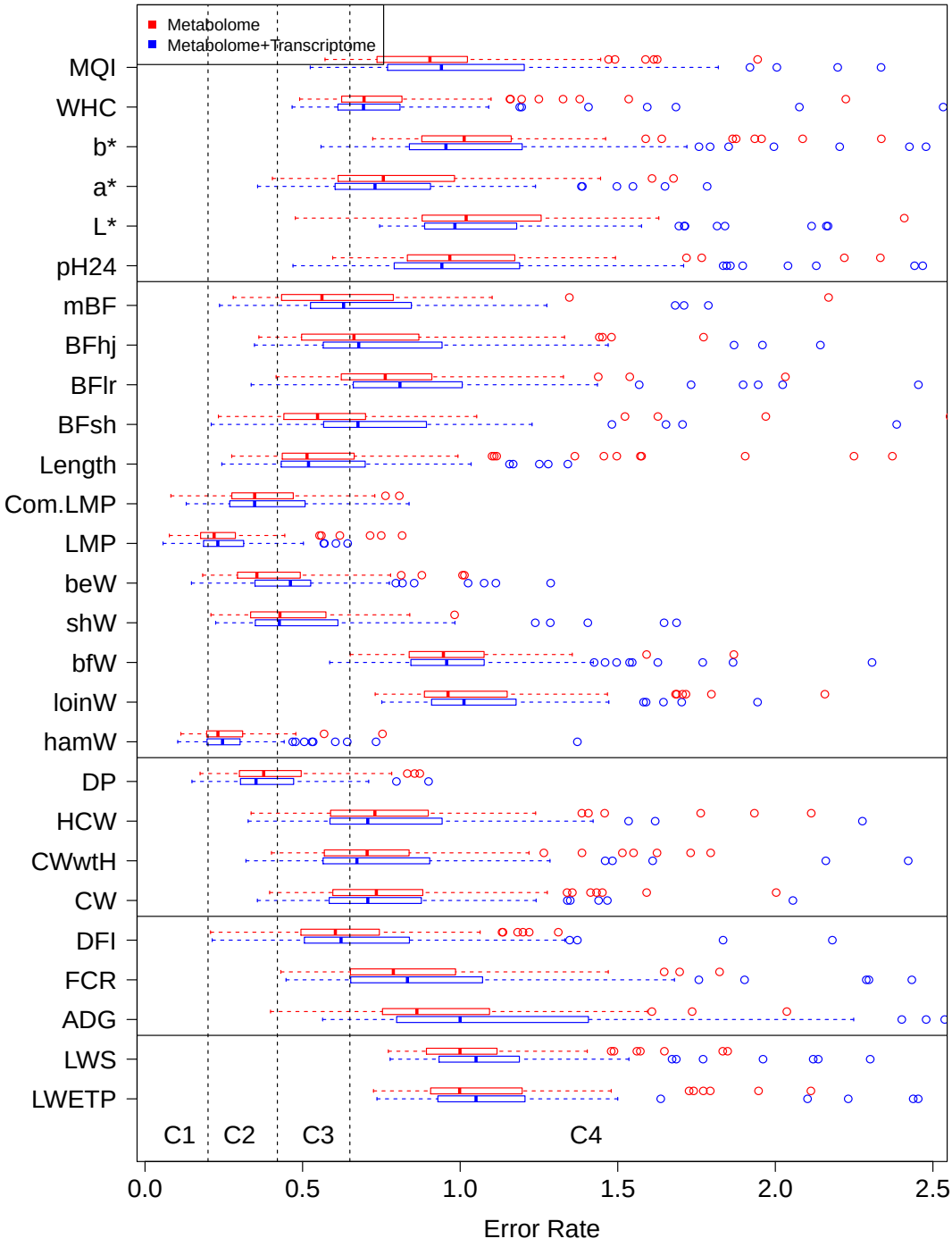


FIGURE 10 – **Prédictions de 27 phénotypes par le métabolome et par la combinaison du métabolome et du transcriptome, avec la méthode Lasso en prenant en compte la race des individus.**

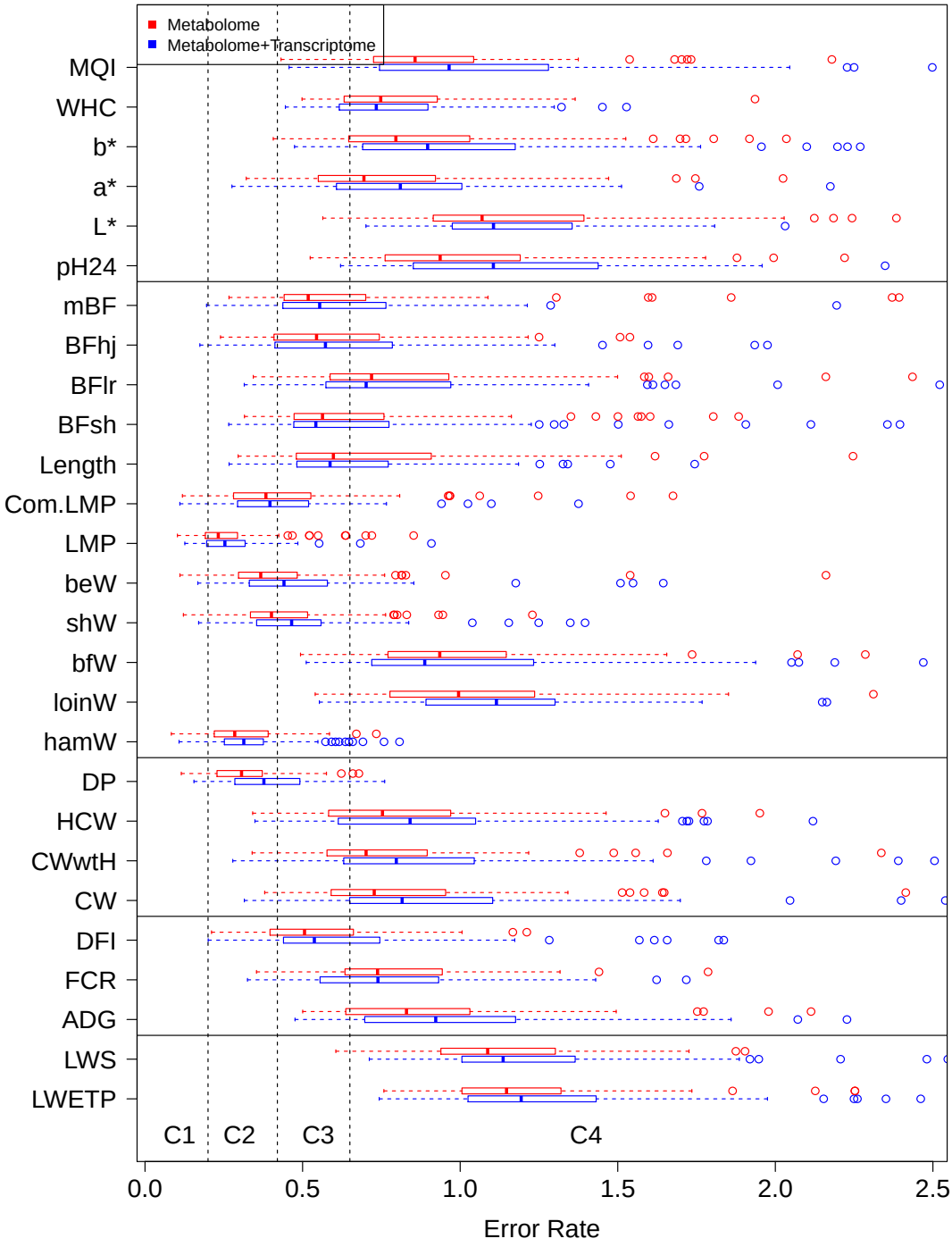


FIGURE 11 – **Prédictions de 27 phénotypes par le métabolome et par la combinaison du métabolome et du transcriptome, avec la méthode Lasso en prenant en compte la race et la bande des individus.**

6 Conclusion

Les objectifs de cette thèse étaient la prédiction de phénotypes d'intérêt économique ainsi que la détermination de marqueurs biologiques permettant d'expliciter un phénotype, cela à partir de données métabolomiques recueillies sur des individus en croissance.

S'affranchissant d'un paramètre de régularisation très important pour la plupart des méthodes classiques, de nouvelles méthodes de sélection de variables ont été développées pour répondre aux problèmes posés, que ce soit dans un modèle linéaire ou dans un modèle linéaire mixte. Dans le premier, les méthodes sont des procédures séquentielles de tests multiples ayant des résultats de puissance non asymptotique dépendants de la force du signal, et de plus, elles fonctionnent en grande dimension ($p > n$). Ces procédures donnent de très bons résultats en simulation et des résultats mitigés sur les données réelles provenant du projet DéLiSus de l'INRA. Les modèles mixtes ont donc été étudiés et un algorithme a été développé pour la sélection d'effets fixes. Cet algorithme est performant en grande dimension et plus rapide que les méthodes existantes puisqu'il n'est pas basé sur l'inversion d'une matrice de taille $n \times n$ à chaque étape du processus de convergence.

Les résultats présents dans ce manuscrit ont permis la mise en évidence de relations entre certains phénotypes de production et le métabolome, les données métabolomiques ayant un pouvoir prédictif différent pour chaque phénotype. On pourrait toutefois attendre de meilleurs résultats si certaines conditions étaient remplies, comme par exemple un plan d'expérience plus équilibré, ou encore une réduction du délai temporel entre la prise de sang et la mesure des phénotypes puisque l'on sait que le métabolome évolue dans le temps et qu'il reflète un instant précis de la vie de l'animal.
De plus, nous avons commencé par un travail sur trois grandes races porcines -Large White type Femelle, Landrace et Pietrain- qui sont celles qui comportaient le plus d'individus, les analyses présentes dans ce manuscrit doivent donc être poursuivies en incluant toutes les races à disposition (au nombre de 8).
Il est à noter que des analyses à deux tableaux ont été envisagées, en considérant tous les phénotypes comme faisant partie d'un même tableau, mais elles n'ont pas été poursuivies puisque le travail s'est rapidement focalisé sur des phénotypes particuliers comme la consommation journalière -DFI- ou le taux de muscle -LMP-. Cependant, des méthodes permettant de considérer deux tableaux existent, comme la PLS (Partial Least Squares) (Wold, 1966), la sPLS (sparse PLS) (Lê Cao et al., 2008) ou l'analyse de co-inertie (Dolédec and Chessel, 1994).

Tout le travail fourni dans ce manuscrit avait pour but de répondre à des questions biologiques précises et appliquées dans le domaine agronomique. Cependant, les méthodes existantes ainsi que les méthodes nouvellement développées peuvent être appliquées dans des champs plus diversifiés de la recherche ou de la science en général. En effet, réussir à expliciter les éléments d'un objet qui prédominent dans la relation entre deux objets quels qu'ils soient est une question très répandue dans le monde de la science.

Références

Bach, F. (2009). Model-consistent sparse estimation through the bootstrap. Technical report, hal-00354771, version 1.

Baraud, Y., Giraud, C., and Huet, S. (2009). Gaussian model selection with an unknown variance. *Ann. Statist*, 37(2) :630–672.

Baraud, Y., Huet, S., and Laurent, B. (2003). Adaptative test of linear hypotheses by model selection. *Ann. Statist.*, 31(1) :225–251.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate : a Practical and Powerful Approach to Multiple Hypothesis Testing. *J. R. Stat. Soc.*, B 57, 289-300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the False Discovery Rate in multiple testing under dependency. *Ann. Statist.*, 29(4) :1165–1188.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37(4) :1705–1732.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) :203–268.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection of fixed and random effects in linear mixed-effects models. *Biometrics*, 66 :1069–1077.

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

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Statist.*, 1 :169–194.

Bunea, F., Wegkamp, M., and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Statist. Plann. Inference*, 136 :4349–4363.

Candes, E. and Tao, T. (2007). The Dantzig selector : Statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6) :2313–2351.

Causeur, D., Friguet, C., Houee-Bigot, M., and Kloareg, M. (2011). Factor Analysis for Multiple Testing (FAMT) : An R-package for Large-Scale Significance Testing under Dependence. *Journal of Statistical Software*, 40(14).

Chen, J. and Chen, Z. (2008). Extended Bayesian Information Criteria for Model Selection with Large Model Spaces. *Biometrika*, 94 :759–771.

Chesneau, C. and Hebiri, M. (2008). Some theoretical results on the Grouped Variables Lasso. *Mathematical Methods of Statistics*, 17(4) :317–326.

Dolédec, S. and Chessel, D. (1994). Co-inertia analysis : an alternative method for studying species-environment relationships. *Freshwater*, 31(3) :277–293.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2) :407–499. With discussion, and a rejoinder by the authors.

Foulley, J. (1997). ECM approaches to heteroskedastic mixed models with constant variance ratios. *Genetics Selection Evolution*, 29 :197–318.

Foulley, J.-L., Delmas, C., and Robert-Granié, C. (2002). Méthodes du maximum de vraisemblance en modèle linéaire mixte. *J. SFdS*, 1-2 :5–52.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, 72 :320–340.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9 :226–252.

Henderson, C. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, pages 10–41.

Henderson, C. (1984). *Applications of linear models in Animal breeding*. University of Guelph, Ont.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptative lasso for sparse high-dimensional regression models. *Stat. Sin.*, 18(4) :1603–1618.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*, 67 :495–503.

Lavergne, C., Martinez, M. J., and Trottier, C. (2008). Empirical model selection in generalized linear mixed effect models. *Comput. Statist.*, 23 :99–109.

Lê Cao, K. A., Boitard, S., and Besse, P. (2011). Sparse PLS discriminant analysis : biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12 :253.

Lê Cao, K. A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7 :Article 35.

McLachlan, J. and Krishnan, T. (2008). *The EM Algorithm and Extensions, second edition*. Wiley-Interscience.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34(3) :1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). *J. R. Stat. Soc. : Series B*, 72 :417–473.

Meng, X.-L. and Rubin, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm : A general Framework. *Biometrika*, 80 :267–278.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58 :545–554.

Rao, C. R. and Wu, Y. H. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2) :369–374.

Rohart, F. (2012). Multiple Hypotheses Testing For Variable Selection.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization. *Scandinavian Journal of Statistics*, 38 :197–214.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist*, 6(2) :461–464.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, B 58(1) :267–288.

Verzelen, N. (2012). Minimax risks for sparse regressions : Ultra-high-dimensional phenomenons . *Electron. J. Statist.*, 6(1) :38–90.

Wainwright, M. (2009). Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ^1-Constrained Quadratic Programming (Lasso). *Information Theory, IEEE Transactions on*, 55 :2183–2202.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.*, 37 :2178–2201.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis.*, page 391–420. p.r. Krishnaiah (Ed.), New York, Academic Press.

Yuan, L. and Lin, Y. (2007). Model selection and estimations in regression with grouped variables. *J. R. Stat. Soc. : Series B*, 68 :49–67.

Zhang, C.-H. and Hunag, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4) :1567–1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7 :2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc. 101*, 101(476) :1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J.R. Statist. Soc.*, B 67(2) :301–320.