Université Évry Val d'Essonne
Laboratoire Statistique et Génome

# Thèse

présentée en première version en vu d'obtenir le grade de Docteur,
spécialité Mathématiques Appliquées

par

Camille Charbonnier

## Inférence de réseaux de régulation génétique à partir de données du transcriptome non indépendamment et identiquement distribuées

---

## Inference of gene regulatory networks from non independently and identically distributed transcriptomic data.

Thèse soutenue le 4 décembre 2012 devant le jury composé de :

| | | | |
|---|---|---|---|
| M. | Christophe Ambroise | Université d'Évry Val d'Essonne | (Directeur) |
| M | Stéphane Canu | INSA Rouen | (Rapporteur) |
| M. | Julien Chiquet | Université d'Évry Val d'Essonne | (Co-directeur) |
| M. | Christophe Giraud | Université Paris 11 | (Examinateur) |
| Mme | Sylvie Huet | INRA | (Rapporteur) |
| Mme | Sophie Lèbre | Université de Strasbourg 1 | (Examinateur) |
| M. | Franck Picard | CNRS | (Examinateur) |

[NNT : 2012EVRY0022]

# Remerciements

Je suis très heureuse de trouver en cette page un espace consacré au remerciement de tous ceux sans qui cette thèse ne serait ni ce quelle est, aujourd'hui, telle que circonscrite dans ce manuscrit, ni ce qu'elle a été, à travers ces trois années d'aventures doctoresques.

En premier lieu, je pense à Christophe et Julien. En deuxième année d'ENSAE, Christophe m'a ouvert les portes d'un monde où les statistiques n'étaient pas seulement un outil d'analyse du monde économique et social, mais une fenêtre ouverte sur les dernières avancées de la recherche en biologie, de quoi garder les yeux pétillants de curiosité et d'émerveillement. Ils m'ont tous les deux accueillis en milieu de master pour définir un sujet de thèse où les développements statistiques côtoieraient les problématiques biologiques. Pendant les trois ans qui ont suivi, leur bienveillance, leur disponibilité et leur confiance m'ont permis de dépasser les doutes et tergiversations d'une thésarde. Je suis particulièrement honorée d'avoir été la première thésarde de Julien, et je mesure tout le dévouement avec lequel il m'a accompagnée pendant ces trois ans. Je les remercie par ailleurs de toutes les opportunités qu'ils m'ont offertes ou incitée à saisir, ainsi que de la liberté qu'ils m'ont laissée de travailler indépendemment avec Fanny et Nicolas. J'espère cultiver à distance les enseignements de cette expérience à leur côté pour m'inspirer tout autant de la sérénité et de l'optimisme de Christophe, que de la persévérance et du dynamisme de Julien. [1]

Je remercie vivement Sylvie Huet et Stéphane Canu, qui m'ont fait l'honneur et la joie de rapporter ce travail. La finesse et la rigueur de leurs relectures ont été essentielles à l'amélioration de ce manuscrit et l'ouverture de nouvelles pistes de réflexion.

Je remercie également Christophe Giraud d'avoir présidé le jury de thèse. Je le remercie, en même temps Sophie Lèbre et Franck Picard, de l'intérêt qu'ils ont tous trois porté à ce travail.

Je pense aussi à tous les membres du Laboratoire Statistique et Génome, pour la convivialité qu'ils ont su y cultiver, convivialité qui met la barre très haute pour tous les prochains/éventuels labos d'accueil. Je ne saurai remercier personnellement chacun sans m'en vouloir ensuite d'avoir omis quelqu'un ou quelqu'anecdote, mais c'est l'heure de se plier à l'exercice : merci à Bernard (~~de savoir fredonner à la volée le concerto pour clarinette de Mozart~~ pour toutes les curiosités mathématiques et autres

[1] J'ajoute entre parenthèse la satisfaction farfelue de me dire que si Julien a accepté d'accompagner dans une bien plus longue aventure un petit Camille, j'ai au moins échoué en trois ans à lui faire détester ce prénom.

que j'ai pu découvrir au détour d'une discussion de couloir, et en souvenir de mes premiers enseignements sur les tests statistiques), Catherine (pour son expertise asymptotique précieuse), à Marie-Luce, Anne-Sophie et Cyril (pour leurs conseils pédagogiques), au trio de bioinfo, Carène, Claudine, Yolande (pour la bonne humeur et les fous rires qu'elles sèment au labo, même si je dois déplorer *a posteriori* la validité de la théorie des trois ans de Carène), Etienne (de nous rappeler qu'on peut aussi chercher avec un papier et un crayon, tout simplement), Pierre (tu as toujours mon numéro si besoin pour garder tes enfants le mercredi), Guillem, Mickaël, Cécile, Michèle (pour prendre tant soin de nous tous), ainsi qu'à Maurice (pour les critiques et ressources cinématographiques, mais pas que [2]). Evidemment, merci à Pierre d'avoir partagé son expérience de thésard $n+2$, à Marine et Matthieu qui ont partagé en simultané les doutes et les joies de ces trois ans. Le RER D aura été témoin de beaucoup de nos vociférations de thésards. Merci à Sarah pour son oreille attentive. Je souhaite tout le courage et tout le meilleur à l'équipe de thésards du labo, Van hanh, Marius, Justin, Sarah, Alia, Morgane,...

J'espère bien vous retrouver à l'avenir en conférence, séminaire ou encore à la campagne si l'envie d'une promenade normande en bord de mer vous gagne (oui, je sais, c'est en dehors et bien loin du périph, mais sympa quand même : Michel et Augustin ne sont pas encore arrivés jusqu'à nous, mais on a des oeufs et légumes frais, sans même acheter de panier bio).

Un grand merci aussi à Fanny et Nicolas, pour tout ce que j'ai appris auprès d'eux, de scientifique ou non, mais surtout pour le plaisir que j'ai à travailler avec eux, comme quoi les retards de la SNCF ont parfois d'excellentes vertus.

Je finis par tous ceux qui n'ont rien à voir de près avec cette thèse, mais sans qui celle-ci ne serait jamais arrivée. Merci à mes parents pour avoir nourri ma curiosité, et pour leur confiance sans faille face à mes choix d'orientation parfois exotiques. Merci à toutes celles et ceux qui ont rendu ces huit années d'études palpitantes, à toutes celles qui ont partagé cette aventure doctoresque en temps réel, à toutes celles qui l'ont accompagnée depuis leur autre univers. Merci en particulier à Isabelle et Anne-Lise d'avoir mis à ma disposition un canapé si douillet par temps de fatigue. Merci pour finir à Sylvain, pour m'accompagner chaque jour dans une bien plus longue aventure.

[2] Pour être tout à fait honnête, je devrais ajouter ici l'inestimable biprofénide de veille de soutenance. Merci !

# Contents

# Introduction

Following the recent outburst of genetic data offered by microarray experiments and whole-genome sequencing, biological phenomena have been screened through many large scale analyses. Genome-wide association studies have investigated potential associations between a phenotype of interest and particular genotypes in terms of single-nucleotide polymorphisms (SNPs). Differential analyses of transcriptomic datasets looked for association between phenotypes and gene expression profiles. Contrary to SNP datasets, which capture the variability of nucleotide sequences at the individual level, transcriptomic or expression datasets aim to measure the variability of gene activity over time and tissues. Transcribed mRNA levels are used as a proxy for the levels of proteins used by the cell at a given combination of time, tissue and environmental condition.

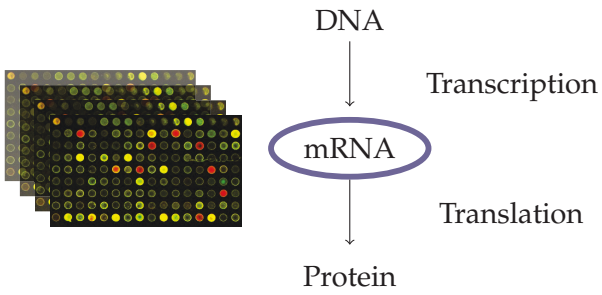



Figure 1 – *Expression data: measurements of mRNA levels used as a proxy for gene activity.*

Despite great breakthroughs, those analyses stumble upon the fact that the penetrance of single candidate alleles often remains very low (Varghese and Easton 2010) or gene signatures suffer from high variability (Ioannidis 2005, Haury et al. 2011a). Among others, one explanation is that they consider each gene independently and miss to take into account their interactions. There is therefore an increasing interest for multivariate approaches, adopting the approach of *systems biology*.

From a mathematical viewpoint, graph theory provides an ideal framework to model biological systems. A graph $\Gamma$ is defined as a couple $(\mathcal{V}, \mathcal{E})$ of vertices and edges. Depending on whether the graph is directed or not, that is to say whether the edges are directed or not, the set $\mathcal{E}$ denotes a set of ordered (resp. unordered) pairs of vertices. Many biological phenomena can be represented under the form of a graph, or network[3]. Roughly speaking, we can distinguish at least three types of well-modeled

[3]In the sequel, we use indifferently the vocabulary of "graph" or "network". The former is more familiar to the mathematical community while the latter is more often used in the biological one.

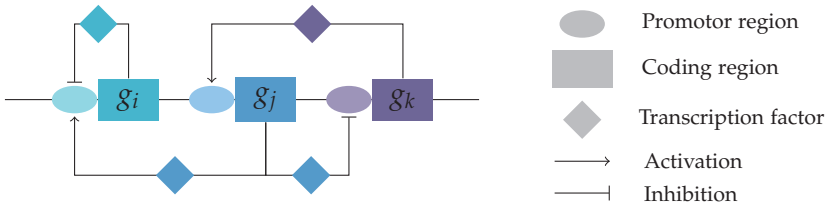



Figure 2 – *Targeted model of regulatory mechanism: some genes code for proteins, called transcription factors, which bind to the promoter region of other genes in order to regulate their activity.*

biological networks: protein-protein interaction (PPI) networks, metabolic pathways, and gene regulatory networks.

PPI networks model how proteins bind to each other. Vertices consist of proteins; edges are added between two proteins when those are known to bind together. PPI networks are a fruitful source of information but there is a huge variation among them depending on the definition used for the binding and the techniques developed in order to identify them. To illustrate the wide variety of PPI sources we could mention, among others, physical experimentations with potential discrepancies between *in sillico* or *in vivo* methods, and computational biology algorithms, based for instance upon phylogeny and homology relationships, 3D structure modeling, or supervised learning techniques.

Metabolic networks compile metabolic pathways, which are the chains of chemical reactions responsible for specific biological functions taking place in a cell. In metabolic networks, an edge links in chain substracts and products of chemical reactions, such that products of one reaction play the role of substracts to the next. Metabolites involved in chemical reactions consist in gene products and transformation of gene products but also in a large range of cofactors found in the environment. More refined metabolic networks also provide information on the enzymes required to catalyse the reaction, which make them quite complex and heterogeneous.

Gene regulatory networks aims to describe the inhibition and activation relationships operated by transcription factors onto genes, as illustrated in Figure 2. As such, each vertex represents at the same time a gene and its protein products as one while edges represent the fact that one of the genes codes for a protein which binds to the promoter region of the other in order to regulate its activity. If gene regulations can be identified individually via biological experimentations like knock-outs, it is now an important statistical challenge to recover those gene regulatory networks on a large scale thanks to expression datasets.

This statistical issue is the main motivation of this thesis. In such a case, our definition of gene regulatory networks is the biological target that we try to model. Yet, the acurate interpretation of what we infer is conditioned by two points: first, by the statistical modeling, which will be detailed and discussed along the manuscript, second by the data on which we work. Indeed, we restrict ourselves to the observation of biological phenomena from the unique point of view of mRNA levels, omiting the multiplicity of regulatory actors and the complexity of regulatory mechanisms themselves.

Various statistical techniques have already been studied to tackle this

issue, among which partial derivative equations and Bayesian dynamic networks. In this thesis we adopt the framework of Gaussian graphical models (GGMs), which combines both assets of multivariate Gaussian distributions and graph theory.

Chapter 1 recalls major recent developments in this area, when the main objective is not to infer the distribution at known graphical structure but to recover the graphical structure itself. In that respect, the leading challenge resides in design proportions, since we face expression datasets where the number of available microarrays is much smaller than the number of genes under study. This so-called *high-dimensional* setting most certainly defines a new paradigm for recent statistical developments, at the opposite end of the usual asymptotic framework which consists in allowing the number of observations to grow to infinity in order to obtain the most accurate estimations. The first challenge of high-dimensional statistics is to even be able to provide an answer in a context where it seems *a priori* impossible using classical methods. Chapter 1 therefore describes the main advances offered by regularized approaches to solve high-dimensional problems. At the root of the high-dimensional paradigm is the notion of *sparsity*, which assumes that the burden of dimension is only apparent: the true size of the problem is actually much lower than it seems, and it suffices to look for a solution in subspaces of low dimension where the problem is solvable. Regularization and shrinkage approaches described in Chapter 1 provide a particularly efficient way of exploring those low dimension subspaces.

In that context, data heterogeneity might be an asset, particularly to improve the quality of the answer when the sample size is low. Chapters 2 and 3 consider two different but possibly complementary definitions of heterogeneous transcriptional data. In both cases, an adequate statistical modeling can alleviate the burden of high-dimension.

Chapter 2 models heterogeneity at the network level, building upon the assumption that biological networks are organized: genes known to participate in the same biological functions are more likely to regulate each other, while some of them coding for transcription factors are much more likely to play the role of hubs in networks. Following the work of Ambroise et al. (2009), Chapter 2 suggests to make use of prior information about the topology of the network in the definition of a *weighted-Lasso* estimator to improve the accuracy and robustness of the identification of regulations.

Chapter 3 models heterogeneity at the observation level, focusing on a recent regularization term called the *cooperative-Lasso* designed to combine observations from distinct but close datasets (Chiquet et al. 2011). Since many transcriptomic experiments are led simultaneously in several *close* conditions, as part of a more general experimental scheme, such as stress experiments, case/control or placebo/treatment studies, a method which allows the information to be shared across conditions without reducing the estimation to a single average network as would be done by meta-analysis is of high-interest. This chapter refers to regularization schemes

which has nourished a lot of research in the machine learning community under the term of multi-task learning.

Finally, Chapter 4 addresses the crucial question of uncertainty in an ongoing work in collaboration with F. Villers and N. Verzelen. Chapter 1 details ways to provide an answer in awkward design sizes. Chapter 2 and 3 define ways to provide an hopefully improved answer. Theory states conditions under which the answer is reliable and consistent, however there is a need to quantify the quality and certainty of the answer on a given dataset. Particularly, if we want the networks we infer to be used appropriately by clinicians who want to improve disease diagnostics and prognostics and maybe eventually identify new drug targets, we need to be able to confirm that differences observed between two networks inferred in distinct conditions are indeed significantly different. Chapter 4 tackles this issue.

Our contribution to this chapter focuses on the adaptation of Verzelen and Villers (2010) with the use of Fisher statistics, the adaptation of higher-criticism, as well as numerical experiments and other practical aspects.

# 1 High-dimensional Gaussian Graphical Models and Gene Regulatory Networks

The objective of this chapter is to clarify the statistical framework adopted in this thesis to model and infer gene regulatory networks. We recall the definition and interpretation of Gaussian graphical models, in their directed and undirected forms. We then draw an introduction to $\ell_1$ regularization and its adaptation to the inference of high-dimensional Gaussian graphical models.

Contents

## 1.1 Introduction

Microarray transcriptomic data epitomize the challenges raised by the statistical modeling of complex biological systems in the recent era of high-dimensional data. The following chapter provides insights into the two fundamental stones combined in this thesis to meet this challenge: the modeling of gene regulatory networks by Gaussian graphical models and $\ell_1$ regularization of high-dimensional problems.

Gaussian graphical models (GGMs) provide a theoretically well defined framework to study gene regulatory networks. Admittedly, our models cannot reflect the complex reality of regulatory mechanisms, but at least we can strictly interpret and control what comes out the data from a statistical point of view. The first section of this chapter attempts at clarifying the structure of conditional dependences exhibited by Gaussian graphical models.

In order to deal with high-dimensional data, regularized approaches have nourished an outstanding research effort in the last decades. Among them, $\ell_1$ regularization lies at the boundary between shrinkage convex estimators and model selection. We devote a second section to the mechanims allowing $\ell_1$ regularization to perform simultaneously estimation and model selection and eventually combine $\ell_1$ regularization with the inference of high-dimensional Gaussian graphical models.

## 1.2 Statistical Modeling of Gene Regulatory Networks

Assume that the vector of expression levels $X = (X_1, \ldots, X_p)^\intercal$ follows a regular multivariate Gaussian distribution with expectation $\mu \in \mathbb{R}^p$ and covariance structure $\Sigma \in \mathbb{S}_+$, where $\mathbb{S}_+$ denotes the set of symmetric positive definite matrices. GGMs provides a graphical representation of the conditional dependence structure between components of $X$. Thorough definitions and properties of GGMs can be found in Whittaker (1990) and Lauritzen (1996). Before going further into GGMs we need to clarify what we mean by conditional dependence structure.

In general, the independence of two components $X_i$ and $X_j$ conditionnally on a remaining set of components $C$ describes the fact that knowing $X_C$, $X_j$ does not bring any supplementary information on $X_i$ that is not already brought by $X_C$, and *vice versa*. If $X$ admits a density with regard to a certain measure $\mu$, this conditional independence means that the joint density $f_{i,j,C}(x_i, x_j, x_C)$ of $(X_i, X_j, X_C)$ factorizes into $f_{i,C}(x_i, x_C) f_{j,C}(x_j, x_C)$, or similarly, the conditional distribution $f_{(X_i,X_j)|X_C}(x_i, x_j; x_C)$ factorizes into $f_{X_i|X_C}(x_i; x_C) f_{X_j,X_C}(x_j; x_C)$.

### 1.2.1 Undirected Gaussian Graphical Models

The crucial point in the definition of conditional independence is the set $C$ on which the conditionning is taken. Different Markov properties illustrated in Figure 1.1 distinguish the conditional independence structure represented by an undirected graph $\Gamma$. Starting from the less stringent

definition towards the strongest assumption, it is interesting to recall the definitions of pairwise Markovian, local Markovian or global Markovian, all with respect to a given graph $\Gamma$. It is easily checked that 1.3 implies 1.2 which in turn implies 1.1.

**Definition 1.1** (Pairwise Markov property) *The random vector $X$ is pairwise Markov with respect to a graph $\Gamma = (\mathcal{V}, \mathcal{E})$ if and only if, for every pair of non-adjacent vertices, i.e. for every pair $(i,j) \in \mathcal{V}^2$ such that $(i,j) \notin \mathcal{E}$, $X_i$ is independent from $X_j$ conditionnaly on all remaining components:*

$$i \nleftrightarrow j \quad \Leftrightarrow \quad X_i \perp X_j | X_{\mathcal{V}\setminus\{i,j\}}.$$

**Definition 1.2** (Local Markov property) *The random vector $X$ is local Markov with respect to a graph $\Gamma = (\mathcal{V}, \mathcal{E})$ if and only if, for every vertex $i$, $X_i$ is independent from all its non-neighbors conditionnaly on its neighbors $ne_\Gamma(i)$:*

$$i \nleftrightarrow j \quad \Leftrightarrow \quad X_i \perp X_j | X_{ne_\Gamma(i)}.$$

**Definition 1.3** (Global Markov property) *The random vector $X$ is global Markov with respect to a graph $\Gamma = (\mathcal{V}, \mathcal{E})$ if and only if, for every three subsets distinct $(I, J, S)$ of vertices, such that $X$ separates $I$ from $J$ in $\Gamma$, for every pair $(i,j) \in I \times J$, $X_i$ is independent from $X_j$ conditionnaly on $X_S$:*

$$i \nleftrightarrow j \quad \Leftrightarrow \quad X_i \perp X_j | X_S.$$

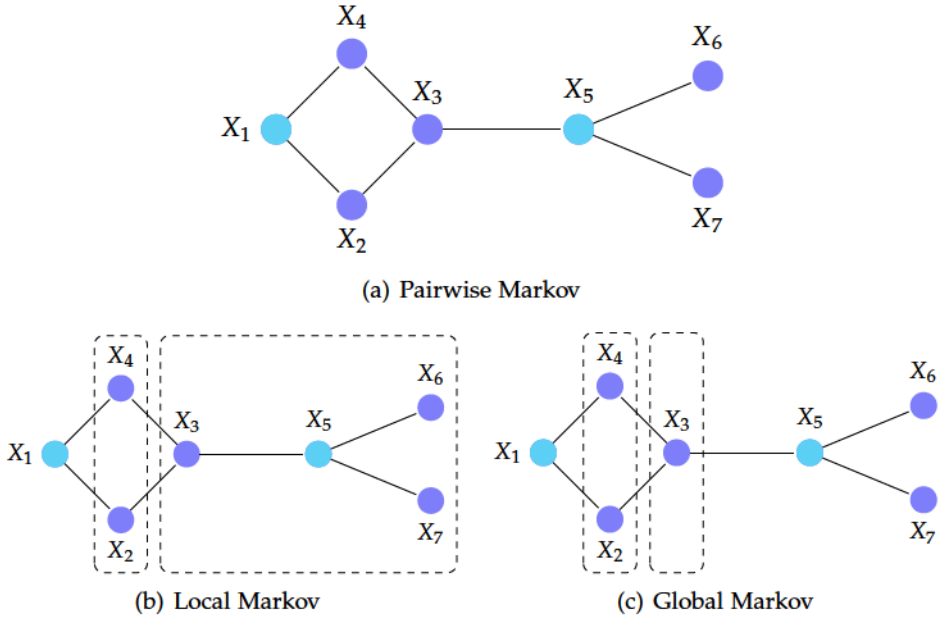
With these definitions in mind, we can now define undirected Gaussian graphical models.

**Definition 1.4** (Gaussian graphical model) *A multivariate Gaussian vector $X = (X_1, \ldots, X_p)$ follows a Gaussian graphical model with respect to a graph $\Gamma = (\mathcal{V}, \mathcal{E})$ if and only if $X$ satisfies the pairwise Markov property with respect to $\Gamma$. For every pair of vertices $(i,j) \in \mathcal{V}^2$ such that $i \neq j$, by:*

$$i \nleftrightarrow j \Leftrightarrow X_i \perp X_j | X_{\mathcal{V}\setminus\{i,j\}}$$

Now, thanks to the Hammersley and Clifford theorem (see Lauritzen (1996)), the existence of a positive and continuous density implies the equivalence, in this particular case, between the three Markov properties. In other words, if $X$ is a GGM with respect to a graph $\Gamma$, it is not only pairwise Markov with respect to $\Gamma$, but also local Markov and global Markov. This equivalence is particularly interesting in terms of interpretation. Indeed, the consequence of the local Markov property is that the best linear prediction of a component $X_i$ is given by its neighbours. Conditional on the set of neighbors $X_{\text{ne}_\Gamma(i)}$, no supplementary information can be extracted from remaining components to improve this prediction. The consequence of the global Markov property is that if two genes $X_i$ and $X_j$ are linked through a certain path of edges, then conditional on any gene on the path, and not only their neighbours, $X_j$ and $X_i$ are independent.

Markov properties clarify the interpretation of the graphical structure provided by the application of GGMs on transcriptomic data. The graphical structure indicates that conditional on the neighbors of a gene, all other genes in the dataset are irrelevant to explain its expression levels. Among a set of correlated genes, we could say that GGMs aims at discovering




(a) Pairwise Markov

(b) Local Markov

(c) Global Markov

Figure 1.1 – *If the vector $X = (X_1, \dots, X_5)$ follows a GGM with respect to the graph in panel (a), then it is equivalently pairwise Markov, local Markov and global Markov with respect to the same graph. Consider for instance the vertices $X_1$ and $X_5$. The pairwise Markov property states that they are independent conditional on all other vertices, that is to say $(X_2, X_3, X_4, X_6, X_7)$. The local Markov property illustrated in panel (b) implies that conditional on either the neighbours of $X_1$ or the neighbours of $X_5$ there is no dependency left between $X_1$ and $X_5$. The consequence of the global Markov property illustrated in panel (c) is that it suffices to condition on any subset of vertices separating $X_1$ from $X_5$ to obtain independence. Particularly, $X_1$ is independent from $X_5$ conditional on $X_3$ alone.*

the direct flow of information from gene to gene that best explains the observed levels of expression.

Thanks to the assumption of Gaussianity, distributions of $X_i$'s conditionnaly on $X_{\setminus i}$ admit very simple expressions. This fundamental result will be at the basis of many further developments. For the sake of clarity, we assume in the following that $X$ is centered. We denote by $A_{i\cdot}$ (resp. $A_{\cdot i}$) the $i$th line (resp. column) of any matrix $A$.

**Proposition 1.1** (Conditional distributions) *Let $X = (X_i)_{i \in \mathcal{V}}$ be a multivariate Gaussian vector with zero mean, invertible covariance matrix $\Sigma = \Theta^{-1}$. The distribution of $X_i$ conditional on $X_{\setminus i}$ is itself a Gaussian distribution with mean $\mu_{i|\setminus i}$ and covariance $V_{i|\setminus i}$:*

$$\mu_{i|\setminus i} = \Sigma_{i\setminus i}\Theta_{\setminus i\setminus i}X_{\setminus i} \qquad V_{i|\setminus i} = \Sigma_{ii} - \Sigma_{i\setminus i}\Theta_{\setminus i\setminus i}\Sigma_{\setminus ii}.$$

Adding the fact that for every $i \neq j$, $\Sigma_{i\cdot}\Theta_{\cdot j} = \Sigma_{i\setminus i}\Theta_{\setminus j} + \Sigma_{ii}\Theta_{ij} = 0$, the conditional distribution of $X_i$ given $X_{\setminus j}$ can be rewritten under the form:

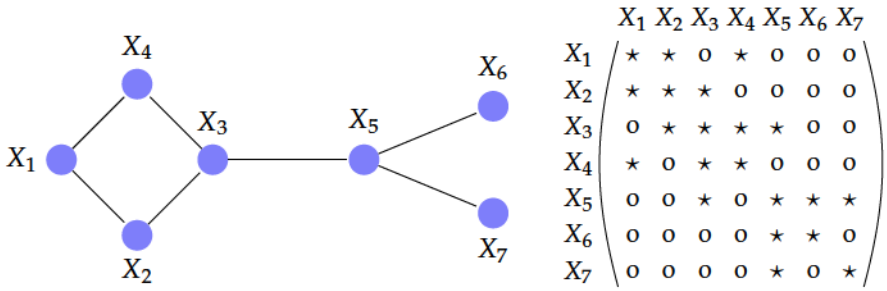
$$X_i | X_{\setminus i} = -\sum_{j \in \mathcal{V}\setminus i} \Theta_{ij}\Theta_{ii}^{-1}X_j + \varepsilon_i,$$

where $\varepsilon_i$ is a centered Gaussian noise with variance $\Sigma_{i,i}$, independent from $X_{\setminus j}$. Under those terms, the neighbours of $X_i$ can be directly read from the set of non zero entries of $\Theta$, which leads to a particularly fruitful characterization of GGMs.

**Proposition 1.2** (Graph of conditional dependencies expressed via the precision matrix) *Let $X = (X_1, \dots, X_p)$ be a multivariate Gaussian vector with zero mean, invertible covariance matrix $\Sigma = \Theta^{-1}$. The graph $\Gamma = (\mathcal{V}, \mathcal{E})$ of conditional dependencies between $X_1, \dots, X_p$ is defined, for every pair of vertices $(i, j) \in \mathcal{V}^2$ such that $i \neq j$, by:*

$$\Theta_{i,j} \neq 0$$

*In other words, if for every vertex $i$, $ne_\Gamma(i)$ describes the set of neighbors of $i$ in the graph $\Gamma$, that is to say the set $\{j \in \mathcal{V}, (i,j) \in \mathcal{E}\}$, the distribution of $X_i$ conditional on $\mathcal{V} \setminus i$ only depends on $X_{ne_\Gamma(i)}$.*




(a) Graph (b) Precision matrix

Figure 1.2 – *The graph of conditional dependencies is characterized by the position of non-zero entries of the precision matrix $\Theta = \Sigma^{-1}$*

This result will be at the center of the following sections and chapters, since statistical methods to recover the graph of conditional dependencies rely on the selection of non-zero entries of the precision matrix.

### 1.2.2 Directed Gaussian Graphical Models

When the interest lies in recovering gene regulations, undirected GGMs provide a somewhat circumsbribed modeling: first, there is no indication of which gene acts as a regulator on the other, second some motifs of particular interest, like retro-active loops, cannot be detected. Directed independence graphs, associated with time-series datasets, in particular directed Gaussian graphical models, provide a fruitful framework to model those. However, the interpretation of edges in terms of Markov properties is a bit trickier.

Although it seems at first sight in complete contradiction with the motivation above, the main constraint that one needs to impose on directed independence graphs is to forbid the presence of cycles, thereby working on directed acyclic graphs (DAGs). We will explain further on how to solve this apparent inconsistency. For now, remark that the presence of cycles would raise problems in terms of factorization of the joint distribution into a chain of conditionnal distributions. Consider a feed-back loop such that $X_1$ regulates $X_2$, which regulates $X_3$, which in turn regulates $X_1$. There would only be a few degenerated cases where we could factorize $f_{(X_1,X_2,X_3)}$ into $f_{X_3|X_2} f_{X_2|X_1} f_{X_1|X_3}$.

A straighforward way to prevent cycles is to provide vertices with a complete ordering $\prec$, so that any edge $i - j$ in the graph can eventually have only one possible direction, such that $i \to j$ if $i \prec j$ and reversely

$i \leftarrow j$ if $j \prec j$. Since a natural ordering is time, notations about directed independence graphs, also called recursive graphs, are traditionaly defined in analogy to genealogic trees, with edges going from parent-nodes to child-nodes. Every node $i$ in $\{1,\dots,p\}$ is endowed with a past $\text{past}(i) = \{1,\dots,i-1\}$, which excludes the present and future nodes $\{i,\dots,p\}$. Given this ordering, edges pointing to $i$ can only come from the past, and edges leaving $i$ can only point to future nodes. For every node $i$, we define its parents $\text{pa}(i)$ as the set of all nodes (in the past), with edges pointing to $i$. As a result, the joint distribution admits a trivial recursive factorization as we hoped:

$$f_{(X_1,\dots,X_p)}(x_1,\dots,x_p) = \prod_{i=1}^{p} f_{X_i|X_{\text{past}(i)}}(x_i\ ;\ x_{\text{past}(i)})$$

**Definition 1.5** (Directed Gaussian graphical model) *A multivariate Gaussian vector $X = (X_1,\dots,X_p)$ follows a directed Gaussian graphical model with respect to a DAG $\Gamma^{\prec} = (\mathcal{V},\mathcal{E}^{\prec})$ if and only if, for every pair of vertices $(i,j) \in \mathcal{V}^2$ such that $i$ belongs to past$(j)$, $X_j$ is independent from $X_i$ conditional on its past (except $i$, naturally):*

$$i \nrightarrow j \Leftrightarrow X_i \perp X_j | X_{past(j)\setminus\{i\}}$$

*The set of nodes from past$(j)$ such that $X_i \not\perp X_j | X_{past(j)\setminus\{i\}}$ is called the set of parents of $j$, pa$(j)$. With this notation, a directed Gaussian graphical model with respect to $\Gamma^{\prec}$ equivalently satisfies the ordered Markov property: for every vertex $i \in \mathcal{V}$,*

$$X_i \perp X_{past(i)\setminus pa(i)} | X_{pa(i)}.$$

It follows that the joint distribution now admits a more refined factorization than the factorization above which corresponded to the saturated graph:

$$f_{(X_1,\dots,X_p)}(x_1,\dots,x_p) = \prod_{i=1}^{p} f_{X_i|X_{\text{pa}(i)}}(x_i\ ;\ x_{\text{pa}(i)})$$

Definition 1.5 echoes the definition of undirected GGMs 1.4 based upon the pairwise Markov property, except that the conditionning on all nodes but $i$ and $j$ is replaced by the conditioning on the past of $j$ (but $i$) only. This definition relies on the existence of an *a priori* ordering of the nodes, which implicitly but strictly governs the direction of edges. As soon as graphs are acyclic, it is always possible to reason the other way round: for every DAG, there exists (though it might be an NP-hard problem to recover it) an ordering $\prec$ of the vertices which is compatible with the DAG, that is to say, for every two vertices such that $i \rightarrow j$, $i \prec j$. The directed local Markov property presents the advantage of not referring to this ordering to provide an interpretation of the edges in terms of conditional dependences.

**Definition 1.6** (Directed local Markov property) *A random vector $X$ satisfies the directed local Markov property with respect to the DAG $\Gamma^{\prec}$ if and onl if for every vertex $i \in$ vertices, $X_i$ is independent from all its non-descendants nd$(i)$ (except its parents) conditional on its parents:*

$$X_i \perp X_{nd(i)\setminus pa(i)} | pa(i)$$

How to obtain the directed global Markov property is less trivial though. Figure 1.3 represents two basic examples chosen by Whittaker (1990) that well illustrate the subtleties of the interpretation of directed edges in terms of conditional dependencies.

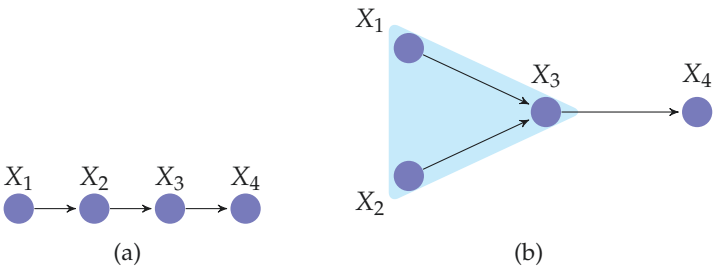



Figure 1.3 – *Two examples of directed graphs. Definition 1.5 associates the two graphs with the following set of conditional independences: in panel (a),* $X_3 \perp X_1|X_2$, $X_4 \perp X_1|X_2, X_3$ *and* $X_4 \perp X_2|X_1, X_3$*; in panel (b),* $X_1 \perp X_2$, $X_4 \perp X_1|X_2, X_3$, $X_4 \perp X_2|X_1, X_3$*. Panel (a) satisfies the Wermuth condition while panel (b) does not because of the motif highlighted by the light blue region.*

Indeed, one would be tempted to interpret directed independence graphs thanks to the useful local or global Markov property. In Figure 1.3, panel (a), it seems natural that $X_4$ should be independent from $X_1$ conditional on either $X_2$ or $X_3$ alone, as it would be the case, would the edges be undirected. The independence conditional on $X_3$, corresponding to the local Markov property, can be obtained by combining $X_4 \perp X_1|X_2, X_3$ and $X_4 \perp X_2|X_1, X_3$. Yet, there is no straighforward way to prove the independence conditional on $X_2$, which would correspond to a global Markov property. Besides, interpreting the directed graph in panel (b) as an undirected one clearly nourishes misleading interpretations. Indeed, in the undirected case, conditioning on $X_3$ leaves $X_1$ and $X_2$ independent while this is absolutely wrong in the directed case.

The difference between the two graphs is the presence of the special motif in panel (b), called Wermuth motif, where the two parents of $X_3$ are independent. To eliminate Wermuth configurations, the idea is to "marry" the parents of the colliders and omit the direction of edges to form the *moral* graph $\tilde{\Gamma}$ associated to the directed graph $\Gamma^{\prec}$. Then the directed independence graph $\Gamma^{\prec}$ shares the same Markov properties as its associated moral graph $\tilde{\Gamma}$. Naturally, when the original directed graph satisfies the Wermuth condition, then it possesses exactly the same Markov properties as its undirected counterpart. The two moral graphs associated with the directed graphs of Figure 1.3 are presented in Figure 1.4.

To state a directed global Markov property without resorting to the moral graph, we need to find out the right definition for a separating set in the case of DAGs. Because of Wermuth configurations, the definition of separating subsets requires a little more definitions than in the undirected case. Along a given trail, let us distinguish collider from non collider nodes: colliders are nodes where edges point to meet. A trail from $i$ to $j$ is said to be blocked by $S$ if and only if either there is a non-collider node within $S$, or if there is a collider node outside $S$ and its ancestor set (the smallest subset containing all the parents of $S$ and all the parents of those

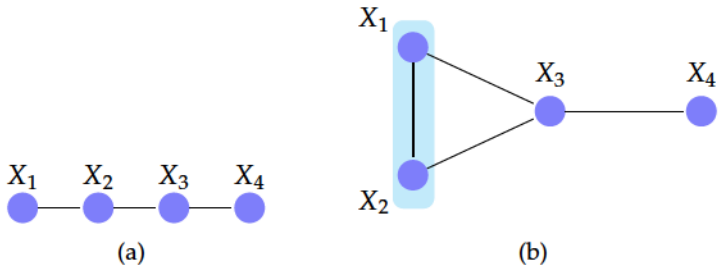



Figure 1.4 – *Moral graphs associated with the graphs of Figure 1.3. Since the directed graph in panel (a) satisfies the Wermuth condition, the associated moral graph is nothing more than its undirected version. The moral graph in panel (b) marries the two parents $X_1$ and $X_2$ of $X_3$, as highlighted by the light blue region. Those moral graphs implies that the original directed graphs also indicate the following conditional independences: in panel (a), $X_3 \perp X_1|X_2$, $X_4 \perp X_1|X_3$, $X_4 \perp X_1|X_2$, and $X_4 \perp X_2|X_3$; in panel (b), $X_4 \perp X_1|X_3$, $X_4 \perp X_2|X_3$ but $X_1 \not\perp X_2|X_3$.*

recusively). A subset $S$ separates two subsets $I$ and $J$ on a DAG $\Gamma^{\prec}$ if and only if $S$ blocks all trails linking $I$ to $J$.

**Definition 1.7** (Directed global Markov property) *A random vector $X$ satisfies the directed global Markov property with respect to the DAG $\Gamma^{\prec}$ if and only if for every pair of vertices $(i, j)$ and subset $S$ separating $i$ from $j$ in $\Gamma^{\prec}$, $X_i$ is independent from $X_j$ conditionnal on $X_S$:*

$$X_i \perp X_j|X_S$$

A typical example of directed GGM is given by autoregressive time-series. Consider a stationnary autoregressive process $X$ of order 1. There exists a parameter $\rho \in ]-1, 1[$ such that for every $t = 1, \dots, T-1$,

$$X_{t+1} = \rho X_t + \varepsilon_{t+1},$$

where $\varepsilon_t$ is the Gaussian white-noise innovation process. The process $\{X_t\}_{t=1}^T$ is a directed GGM with respect to the graph with edges $t \to t+1$ for every $t = 1, \dots, T-1$. Since they admit the same moral graph, it is also a directed GGM with respect to the graph with reversed edges $t+1 \to t$. If the former is in adequation with the natural ordering of time, both are equivalent in terms of Markov properties.

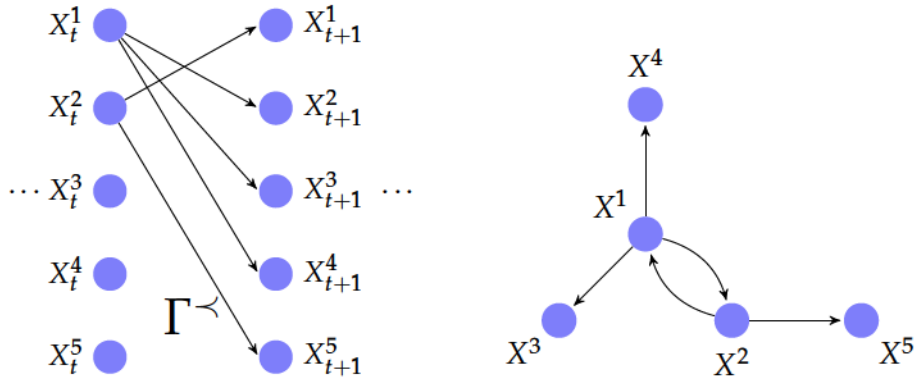
Chapter 2 focuses on the inference of a directed GGM associated with a stationnary autoregressive random vector $X = (X^1, \dots, X^p)$ of order one. There exists a matrix $A$, with eigenvalues smaller than 1 in absolute value, such that for every $t = 1, \dots, T-1$

$$\begin{pmatrix} X_{t+1}^1 \\ X_{t+1}^2 \\ \vdots \\ X_{t+1}^p \end{pmatrix} = A \begin{pmatrix} X_t^1 \\ X_t^2 \\ \vdots \\ X_t^p \end{pmatrix} + \begin{pmatrix} \varepsilon_{t+1}^1 \\ \varepsilon_{t+1}^2 \\ \vdots \\ \varepsilon_{t+1}^p \end{pmatrix}.$$

Under the assumptions of Chapter 2 on the Gaussian white noise and on the absence of within time effects, the process $\{X_t\}_{t=1}^T$ is naturally a directed GGM with respect to the full graph with all edges $X_t^i \to X_{t+1}^j$ for every $t = 1, \dots, T-1$, and every pair of nodes $(i, j)$. However, this

full graph is no more useful than the saturated graph in the undirected case. Chapter 2 will detail the recovery of a minimal directed GGM, representing the minimal across-time conditional dependency structure among nodes.

To finish with, let us come back to the original paradox and explain how this longitudinal representation allows the modeling of feed-back loops despite the absence of cycles. Indeed, edges can either all point from time $t$ to time $t+1$, or be all reversed as we saw in the previous example, hence the absence of cycles. However, if we omit the time-lapse and abusively merge nodes $\{X_1^i, \dots, X_T^i\}$ corresponding to the same component $i$ into one, cycles can appear, as illustrated on Figure 1.5.




(a) Directed GGM (b) Condensed directed GGM

Figure 1.5 – *An extract from a directed GGM representing an order 1 autoregressive random vector, and its condensed representation, omitting the time-lapse over which the correlations take place. Cycles appear in the condensed version while the actual directed GGM representation is acyclic.*

## 1.3 Statistical Inference of High-Dimensional GGMs via $\ell_1$ Regularization

Since we deal with high-dimensional data, it is worth taking a short detour and recall what solutions have been developped in the last ten or twenty years to tackle high-dimensional linear regression, when the number of variables is far larger than the sample size. Next sections will explain how to adapt those preliminary results to the inference of GGMs. For the moment, consider that we observe a size-$n$ sample $(\mathbf{y}, \mathbf{X}) \in \mathbb{R}^n \times \mathcal{M}_{n,p}$ of the following Gaussian linear regression model:

$$Y = X\boldsymbol{\beta}^\star + \varepsilon$$

where $\varepsilon$ is a Gaussian white noise with variance $\sigma^2$.

In high-dimensional settings, the ordinary least square estimator (OLS) is not defined. Assuming that the true parameter $\boldsymbol{\beta}^\star$ lies in a subspace of smaller dimension, one way to provide an answer to the minimization of the quadratic risk is namely to restrict the estimator to lie in a subspace of reduced dimension where a solution exists thanks the addition of penalty terms. Instead of solving the usual least square problem, solve for instance

a problem of the form of (1.1).

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 + \lambda_n \text{ pen}(\boldsymbol{\beta}). \tag{1.1}$$

In Problem 1.1, $\lambda_n$ tunes the amount of shrinkage imposed on $\hat{\boldsymbol{\beta}}$. Ridge regression is a particular case of regularized least square problem, with $\text{pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$. As another particular case, $\ell_1$ regularization, presented in Equation (1.2), has drawn much of research attention since the publications of Donoho and coauthors under the terms of *basis pursuit* (Chen and Donoho 1994) and Tibshirani under the denomination of the *Lasso* for Least Absolute Shrinkage and Selection Operator (Tibshirani 1996).

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 + \lambda_n \|\boldsymbol{\beta}\|_1. \tag{1.2}$$

The specificity of $\ell_1$ regularization compared to $\ell_2$ or any other $\ell_p$ regularization with $p > 1$ is that it is the convex relaxation of the $\ell_0$ pseudo-norm regularization presented in Equation (1.3).

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 + \lambda_n \|\boldsymbol{\beta}\|_0. \tag{1.3}$$

The $\ell_0$ pseudo-norm counts the number of non-zero components of $\boldsymbol{\beta}$. Varying the amount on regularization $\lambda_n$, Problem 1.3 is equivalent to looking for the best linear model with only $k$ variables among $p$:

$$\begin{aligned} \hat{\boldsymbol{\beta}} = &\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 \\ &\text{s.t. } \|\boldsymbol{\beta}\|_0 \leq k. \end{aligned} \tag{1.4}$$

Solving Problem 1.3 or equivalently 1.4 is actually the number of variables is small. With specific choices of $\lambda_n$, respectively $\lambda_n(\text{AIC}) = 2/n$ and $\lambda_n(\text{BIC}) = 2\log(n)/n$, Problem 1.3 boils down to AIC and BIC criteria under assumption of known variance, which can be efficiently combined to forward, backward or forward-backward algorithm to perform model selection. When $p$ grows, the number of models to investigate grows as $2^p$ and the exhaustive search becomes impossible. On the contrary, Problem 1.2 is convex and benefits from efficient convex optimization algorithms, thereby solving in one step both problems of estimation and model selection. The following of our thesis will be based on variations upon such $\ell_1$ regularized problems.

In low as in high-dimension, at least four types of problems can be addressed to measure the quality of those estimators:

P1: *prediction* of $y$, in which case $\boldsymbol{\beta}^\star$ is only the focus of attention as a black box leading to $y$. Performances of the estimator $\hat{\boldsymbol{\beta}}$ are measured in terms of a distance between the optimal linear predictor of $y$ given $X$, $X\hat{\boldsymbol{\beta}}$, and its estimation $X\hat{\boldsymbol{\beta}}^\star$.

P2: *estimation* of $\boldsymbol{\beta}^\star$, also known as *inverse problem* in which case $\boldsymbol{\beta}^\star$ is in itself the center of attention. The objective is to dissect the black box and understand its mechanisms. Performances of $\hat{\boldsymbol{\beta}}$ are measured in terms of a distance between $\boldsymbol{\beta}^\star$ and $\hat{\boldsymbol{\beta}}$. When the sample size is smaller than the number of variables, this problem is particularly ill-posed, since multiple $\hat{\boldsymbol{\beta}}$'s can lead to the same prediction $X\hat{\boldsymbol{\beta}}$.

P3: *selection* of relevant components of $\boldsymbol{\beta}^\star$, also referred to as support recovery, which is essential to provide interpretability to high-dimensional models. This problem is more demanding than Problem [P2], since including small false-positive coefficients might not impede the success of Problem [P2] while severely deteriorate the success of Problem [P3]. Yet, two-step thresholding approaches can adapt procedures successful in Problem [P2] to answer Problem [P3].

P4: *detection* answers the question of whether there is any significant signal in $\boldsymbol{\beta}^\star$. Contrary to Problem [P3], which identifies where exactly the signal is, Problem [P4] only looks for the presence of any signal at all. In terms of hypothesis testing, while Problem [P3] would amount to testing for each single hypotheses $\mathcal{H}_{0,i} : \boldsymbol{\beta}_i^\star = 0$, Problem [P4] amounts to testing the global hypothesis that $\boldsymbol{\beta}^\star = 0$. In high-dimension, it is sometimes much more realistic to treat Problem [P4] than Problem [P3].

Theoretical properties of $\ell_1$ regularization have been studied in the light of those four issues: necessary conditions and achievable results differ. In the following paragraphs, we provide important insights into the $\ell_1$ norm, mainly in comparison with $\ell_0$ and $\ell_2$ regularizations, and recall the main necessary conditions to tackle Problems [P1], [P2] and [P3]. Chapter 2 and 3 try to improve the quality of answers to Problems [P2] and [P3] in the case of Gaussian graphical models. Chapter 4 tackles Problem [P4] in a two-sample framework, where the objective is to detect the presence of heterogeneity between two high-dimensional linear regressions.

### 1.3.1 High-dimensional variable selection via the Lasso

As the convex relaxtion of $\ell_0$ regularization the Lasso lies at the boundary between non-convex $\ell_\gamma$, $0 < \gamma < 1$, regularized problems offering model selection properties and $\ell_\gamma$, $\gamma > 1$, regularized problems which have the advantage of being convex but do not induce sparsity. The range of $\ell_\gamma$ regularizations define the family of Bridge estimators, for $\gamma > 0$:

$$\hat{\boldsymbol{\beta}}_\gamma = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 + \lambda_n \|\boldsymbol{\beta}\|_\gamma^\gamma. \tag{1.5}$$

It is instructive to have a look at the Lasso from this perspective and compare it to other Bridge estimators, particularly the well-known Ridge regression and model selection (1.3) as the limit of $\ell_\gamma$ regularization for $\gamma$ tending to 0.

**How Does $\ell_1$ Regularization Act as a Selection and Estimation Tool?**

Under the light of geometrical, asymptotic and analytic arguments, we want to shed light on why $\ell_1$ is so different from other convex Bridge estimators, with $\gamma > 1$, explaining why $\ell_1$ is the only convex Bridge estimators capable of inducing sparsity. These three frameworks also clarify its limitations as a biased estimation tool.

**The Geometric Point of View: Singularities Induce Sparsity.** The first way to understand how the Lasso can act as a selection tool is to look at it from a geometric argument from convex optimization theory. Indeed, while the least square criterion is differentiable everywhere on $\mathbb{R}^p$, the $\ell_1$ norm is not differentiable at 0. The consequence is that instead of a unique derivative satisfying optimal conditions at points which cross the axes, there is a non-degenerate convex range of possible subgradients which correspond to the same optimal sparse point.

The simplest way to see it is to consider the constrained formulation of Problem 1.5:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 \quad \text{s.t. } \|\boldsymbol{\beta}\|_\gamma^\gamma \leq t. \tag{1.6}$$

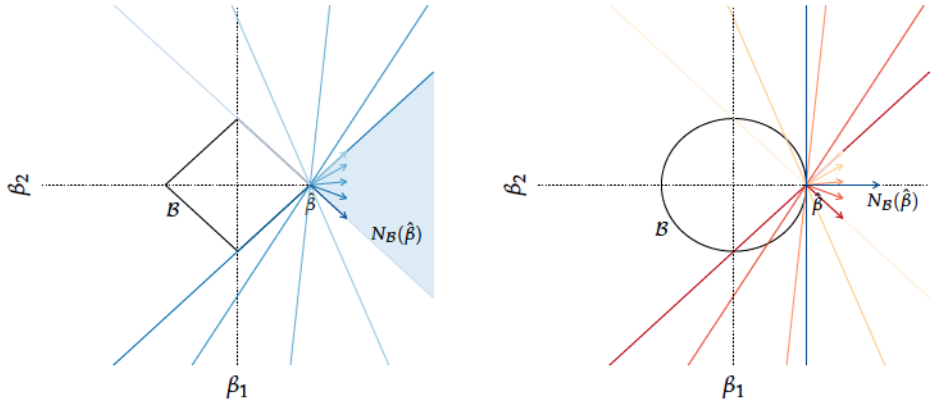
Problem 1.5 is equivalent to the Lagrangian formulation of the constrained formulation, and for every $\lambda_n$ in 1.5, there exists a $t$ in (1.6) such that both problems share the same solution.

First-order optimality conditions for Problem 1.6 state that a point $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is optimal if and only if the least square derivative $-\nabla\ell(\hat{\boldsymbol{\beta}}; \mathbf{y}, \mathbf{X}) = -\mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ defines a supporting hyperplane to the feasible set at $\hat{\boldsymbol{\beta}}$. In other words, the opposite of the least square derivative must belong to the normal cone to the feasible set at $\hat{\boldsymbol{\beta}}$, where the normal cone to a convex set $C$ at point $x_0$ is defined by $\{y \in \mathbb{R}^p, \langle y, x - x_0\rangle \leq 0, \forall x \in C\}$. Thereby, for every $\boldsymbol{\beta}$ in the feasible set, the least square derivative must satisfy

$$\langle \mathbf{X}^\intercal\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rangle \geq 0.$$

In the case of $\ell_\gamma$ norms or pseudo-norms, the feasible set is nothing more than the corresponding ball of radius $t^{1/\gamma}$ in $\mathbb{R}^p$. Figure 1.6 pictures unit balls $\mathbb{R}^2$ for $\ell_1$ and $\ell_2$ balls, along with their normal cones at $(1, 0)$. If we think of the least square derivative as a continuous random variable (as function of the error term $\varepsilon$), then it will almost-never fall into the normal cone to the $\ell_2$ ball at $(1, 0)$, which is degenerated into a single half-line of zero Lebesgue mass. On the contrary, there is non negligeable probability for it to fall into the normal cone to the $\ell_1$ ball at $(1, 0)$, thanks to the singularity.

In other words, contrary to the $\ell_2$ norm which is differentiable on $\mathbb{R}^p$, the $\ell_1$ norm favors the selection of its points of singularities, which are interestingly located on the axis, thereby shrunking some coefficients to exactly 0.




(a) Optimal point $(1,0)$ on the $\ell_1$ ball (b) Optimal point $(1,0)$ on the $\ell_2$ ball

Figure 1.6 – *Optimality conditions for the sparse point* $(1,0)$ *for constrained problems* (1.6) *for the Lasso* ($\gamma = 1$) *and Ridge regression* ($\gamma = 2$)

**The Asymptotic Point of View.** An interesting comparison between $\ell_0$, $\ell_1$ and $\ell_2$ regularizers appears in Knight and Fu (2000). Although the analysis is led in the classical $n > p$ setting, the comparison of $\ell_1$ to $\ell_\gamma$ with $\gamma < 1$ and $\gamma > 1$ limiting distributions is thought-provoking. In this framework, we can assume that $\mathbf{X}^\intercal\mathbf{X}/n$ converges to a positive definite matrix $\Psi$ when $n$ tends to infinity. Then $\mathbf{X}^\intercal\varepsilon/n$ admits a centered Gaussian limit distribution $W$, with variance $\sigma^2\Psi$. Knight and Fu (2000) suggest the amount of regularization follows a $1/\sqrt{n}$ decay rate, such that $\lambda_n\sqrt{n} \to \lambda_0$, then for every $\gamma > 0$, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)$ converges in distribution to $\arg\min(V^\gamma)$ where $V^\gamma$ is defined by:

$$V^\gamma(\theta) = -\theta^\intercal W + \frac{1}{2}\theta^\intercal\Psi\theta + \lambda_0\mathbf{pen}_\gamma(\theta;\boldsymbol{\beta}^\star).$$

All limiting distributions diverge from the OLS limiting distribution by the limiting penalty term $\mathbf{pen}_\gamma$, function of the true signal $\boldsymbol{\beta}^\star$, which distinguishes the three cases $\gamma < 1$, $\gamma = 1$,$\gamma > 1$.

Start with the Lasso, that is to say $\gamma = 1$.

$$\mathbf{pen}_1(\theta;\boldsymbol{\beta}^\star) = \sum_{j=1}^{p} \theta_j \mathrm{sign}(\boldsymbol{\beta}_j^\star)\mathbf{1}_{\{\boldsymbol{\beta}_j^\star \neq 0\}} + |\theta_j|\mathbf{1}_{\{\boldsymbol{\beta}_j^\star = 0\}}.$$

At the limit, the singularity brought by the absolute value is restrained in the limiting objective function to the positions of true zeros. There is no singularity left at truely relevant coefficients, but a bias remains to be paid. In comparison, Bridge estimators like Ridge regression, with $\gamma > 1$ do not exhibit this singularity, as expected from the geometric point of view. Another point of divergence with the Lasso is that the amount of penalty increases with the magnitude of the coefficients, which may lead to unacceptably large bias on most relevant and significant coefficients.

$$\mathbf{pen}_{\gamma>1}(\theta;\boldsymbol{\beta}^\star) = \sum_{j=1}^{p} \theta_j \mathrm{sign}(\boldsymbol{\beta}_j^\star)|\boldsymbol{\beta}_j^\star|^{\gamma-1}$$

The paper underlines the interesting properties of $\ell_\gamma$ pseudo-norms with $\gamma < 1$, which can switch off unrelevant covariates while still estimating true coefficients at the usual $\sqrt{n}$ rate without any bias, since the limiting penalty term remains only active on true zeros.

$$\mathbf{pen}_{\gamma<1}(\theta; \boldsymbol{\beta}^\star) = \sum_{j=1}^{p} |\theta_j| \mathbf{1}_{\{\boldsymbol{\beta}_j^\star = 0\}}.$$

The major inconvenient of $\ell_\gamma$ regularizations with $\gamma < 1$ is that they are non longer convex, loosing all efficient tools of convex optimization, both in terms of computing time and guarantees of converging to a global optimum.

**Bridge Estimators as Thresholding Operators.** The phenomenon observed in limiting distribution is also well illustrated if we express bridge estimators in function of the OLS estimator under an orthonormal design. In this framework, bridge estimators with $\gamma \geq 1$ take the expression of simple thresholding operators, component by component, called proximal operators in the optimization community.

Indeed, considering an orthonormal design such that $\mathbf{X}^\intercal \mathbf{X} = I_p$, the expression of the OLS estimator $\hat{\boldsymbol{\beta}}^{\text{ols}}$ reduces to the product $\mathbf{X}^\intercal \mathbf{y}$, while the opposite least square derivative at $\boldsymbol{\beta}$ simplifies into the difference $\hat{\boldsymbol{\beta}}^{\text{ols}} - \hat{\boldsymbol{\beta}}$. As a result, we can rewrite first-order optimality conditions of the Lasso in terms of $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ and $\hat{\boldsymbol{\beta}}^{\text{ols}}$:

$$\begin{cases} |\hat{\boldsymbol{\beta}}_j^{\text{ols}} - \hat{\boldsymbol{\beta}}_j^{\text{lasso}}| \leq \lambda_n & \text{if } \hat{\boldsymbol{\beta}}_j^{\text{lasso}} = 0, \\ \hat{\boldsymbol{\beta}}_j^{\text{ols}} = \hat{\boldsymbol{\beta}}_j^{\text{lasso}} (1 + \lambda_n |\hat{\boldsymbol{\beta}}_j^{\text{lasso}}|^{-1}) & \text{if } \hat{\boldsymbol{\beta}}_j^{\text{lasso}} \neq 0. \end{cases}$$

Inverse the second equality to $\hat{\boldsymbol{\beta}}_j^{\text{lasso}} = \hat{\boldsymbol{\beta}}^{\text{ols}} - \lambda_n \text{sign}(\hat{\boldsymbol{\beta}}^{\text{ols}})$ and combine it with the constraint of the first inequality to obtain a closed form for $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ as a function of the OLS estimator $\hat{\boldsymbol{\beta}}^{\text{ols}}$:

$$\hat{\boldsymbol{\beta}}_j^{\text{lasso}} = \left(1 - \frac{\lambda_n}{|\hat{\boldsymbol{\beta}}_j^{\text{ols}}|}\right)^+ \hat{\boldsymbol{\beta}}_j^{\text{ols}}$$

In other words, under orthonormal settings, the Lasso operates as a soft-thresholding operator on each covariate independently from the others, substracting $\lambda_n$ to all coefficients (adding for negative coefficients), and switching them off as soon as the absolute value of the OLS estimator goes below the thresholding value of $\lambda_n$.

To obtain the proximal operator related to Ridge regression, rewrite the objective function into

$$\frac{1}{2} \left[ \mathbf{y}^\intercal \mathbf{y} - 2\mathbf{y}^\intercal \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\beta}^\intercal (\mathbf{X}^\intercal \mathbf{X} + 2\lambda_n I_p) \boldsymbol{\beta} \right].$$

As a result of the orthonormality assumption, the first-order optimality condition leads to the following proximal Ridge operator, which shrinks

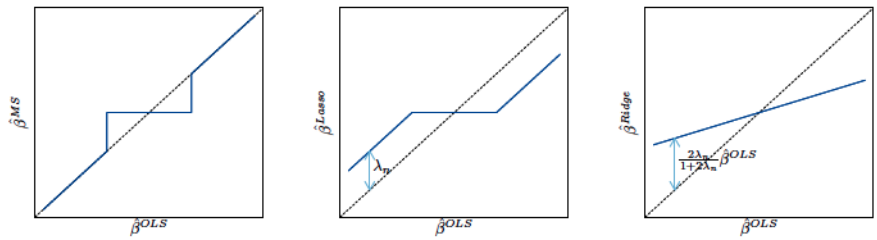
all OLS coefficients by a factor $1/(1+2\lambda_n)$:

$$\hat{\boldsymbol{\beta}}_j^{\text{ridge}} = \frac{1}{1+2\lambda_n}\hat{\boldsymbol{\beta}}_j^{\text{ols}}.$$

The soft-thresholding operator corresponding to the Lasso and the shrinkage operator of Ridge regresssion must be compared to the model selection operated on the basis of a BIC or AIC criterion, or even univariate $t$-test thresholding (along with a correction for multiple testing). Those model selection (MS) approaches correspond to the $\ell_0$ regularized problem and a hard-thresholding operator, so that there would exist a $\lambda_n$ such that:

$$\hat{\boldsymbol{\beta}}_j^{\text{MS}} = \mathbf{1}_{\left\{|\hat{\boldsymbol{\beta}}_j^{\text{ols}}| \geq \lambda_n\right\}}\hat{\boldsymbol{\beta}}_j^{\text{ols}}.$$

All three proximal operators are represented as a function of the OLS estimator in Figure 1.7. As already exhibited by the asymptotic and geometric analyses, model selection seems like an ideal target operator, which identifies a restricted subset of relevant covariates, estimating without bias the coefficient, but remains computionally too demanding for high-dimensional datasets. On the opposite end of the spectrum, Ridge regression is realistic but only acts as a shrinkage estimator, reducing the dimension of the space in which the solution lies without reducing the model size. Not only the final model is too complex to interpret in high-dimention, but the larger the coefficients, the larger the bias. As a compromise between an intractable solution under $\ell_0$ regularization and an $\ell_2$-regularized solution which lacks interpretability, $\ell_1$ regularization offers a realistic approach satisfying both needs for a truely reduced model dimension and a reasonably biased estimation.




(a) Hard-thresholding $\ell_0$ operator (b) Soft-thresholding $\ell_1$ operator (c) Shrinkage $\ell_2$ operator

Figure 1.7 – *Proximal operators corresponding to $\ell_0$ (Model selection), $\ell_1$ (Lasso) and $\ell_2$ (Ridge) regularizations.*

Now that we have exposed the reasons why $\ell_1$ regularization can provide an interesting *shrinkage and selection operator* as Tibhirani put it, and how this selection phenomenon can occur, it is time to raise the question of how *good* is $\ell_1$ regularization in terms of prediction, estimation and selection in the linear regression setting, corresponding to above Problems [P1], [P2], [P3]. For simplicity, we will refer to $\ell_1$ regularization in this setting as the Lasso.

An exhaustive summary of the various assumptions required to guarantee estimation and selection properties of the Lasso in the noiseless case

is given by van de Geer (2009). Figure 1.8 provides a simplified version of Figure 1 in van de Geer (2009).

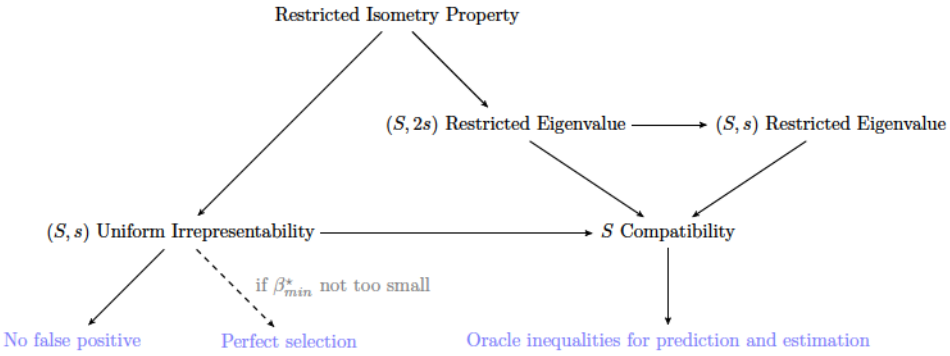



Figure 1.8 – *Summary of causal links between main assumptions required to prove estimation and selection properties of the Lasso, in a simplified representation of Figure 1, reference van de Geer (2009).*

Figure 1.8 highlights the distinction between irrepresentability conditions required for selection consistency and lighter restricted eigenvalue assumptions required for estimation and prediction oracle inequalities. The former has notably been proved necessary for selection properties, and the latter, in its compatibility formulation is possibly the weaker assumption that can be required to obtain at least estimation and prediction consistencies. The next two sections will therefore be devoted to the analysis of those two assumptions. The restricted isometry property is also one of the main assumptions usually used to prove consistency results, but we will not dwell on that one since both previous assumptions are weaker. Besides, these assumptions will be at the basis of the assumptions derived for the cooperative-Lasso in Chapter 3.

### When Does the Lasso Perform Well as a Selection Operator?

The irrepresentable condition, also known as mutual incoherence condition in the community of signal processing, appears simultaneously in a large body of work as a sufficient and necessary condition for selection properties of $\ell_1$ regularized least squares (Zhao and Yu (2006) in statistics, Donoho et al. (2006) and Tropp (2006) in the field of signal processing, while Meinshausen and Bühlmann (2006) defines the equivalent assumption of neighborhood stability). Even though each of these results differ, the main assumption remains the same in both its deterministic design and Gaussian random design forms. Denote by $\mathcal{S}$ the subset of relevant covariates, $\mathcal{S}^c$ its complementary subset.

**Definition 1.8** (Irrepresentable condition for the Lasso under deterministic design) *Consider a fixed design stored in a $n \times p$ matrix $\mathbf{X}$. There exists $\mu > 0$ such that:*

$$\|\mathbf{X}_{\mathcal{S}^c}^\intercal \mathbf{X}_{\mathcal{S}} (\mathbf{X}_{\mathcal{S}}^\intercal \mathbf{X}_{\mathcal{S}})^{-1} sign(\boldsymbol{\beta}^\star_{\mathcal{S}})\|_\infty \leq 1 - \mu. \tag{1.7}$$

**Definition 1.9** (Irrepresentable condition for the Lasso under Gaussian random design) *Consider a Gaussian random design such that each row of the $n \times p$ design matrix $\mathbf{X}$ follows a centered Gaussian distribution with covariance matrix $\Psi$. There exists $\mu > 0$ such that:*

$$\|\Psi_{\mathcal{S}^c\mathcal{S}} (\Psi_{\mathcal{S}\mathcal{S}})^{-1} sign(\boldsymbol{\beta}^\star_{\mathcal{S}})\|_\infty \leq 1 - \mu. \tag{1.8}$$

Parameter $\mu$ is sometimes referred to as the incoherence parameter of exact recovery coefficient. This condition stems from the primal-dual witness construction clearly formulated in Wainwright (2009a) used to prove selection properties of the Lasso, in particular to prove that no irrelevant covariate can be included in the model on top of relevant covariates. Technically, it appears in a three-step reasonning:

1. Infer an oracle estimator $\tilde{\boldsymbol{\beta}}_S$ restricted to the true support $S = S(\boldsymbol{\beta}^\star)$ and complete the estimator by zeros outside the true support, so that this oracle estimator is built to satisfy exact support recovery;
2. Exhibit the subgradient $z$ associated to this oracle $\tilde{\boldsymbol{\beta}}$;
3. Exhibit the (dual feasibility) constraints required on $z$ so that the primal-dual pair $(\tilde{\boldsymbol{\beta}}, z)$ is optimal for the original unconstrained problem, either asymptotically or with large probability.

Conditions 1.7 and 1.8 ensures that, conditional on the inclusion of relevant covariates and non-inclusion of irrelevant ones, the subgradient satifies dual feasibility constraints.

Quite intuitively, these conditions measure in terms of correlation how close irrelevant covariates are to relevant covariates, so that least squares could be misguided into including those irrelevant covariates, hence the regression term of irrelevant covariates onto relevant ones $(\mathbf{X}_S^\intercal \mathbf{X}_S)^{-1}\mathbf{X}_S^\intercal \mathbf{X}_{S^c}$. More precisely, the irrepresentable condition takes the scalar product of this regression term with the true signed support. Indeed, a high-correlation between relevant and irrelevant covariates only presents a risk if it is of the same sign as the true coefficient. Figure 1.9 represent four different situations, two of which satisfy the irrepresentable condition, two others which do not.

The first main results based upon the irrepresentable condition require an asymptotic framework. Wainwright (2009a) introduces a probabilistic approach which allows to work at fixed $n$.

**When Does the Lasso Perform Well as an Estimation or Prediction Operator?**

The irrepresentable condition is quite strong, it is therefore of interest to understand what other possibly good properties could the Lasso demonstrate under weaker conditions. Sparsity oracle inequalities actually show that the Lasso can adapt itself to the true sparsity level in order to perform at minimax rates up to a logarithmic factor in terms of estimation and prediction, under weaker conditions called restricted eigenvalue conditions.

**Definition 1.10** (Restricted Eigenvalue assumption) *Consider a given amount of sparsity $s \leq p$. There exists $\kappa(s) > 0$ such that:*

$$\min_{S \subseteq \{1,\dots,p\}, |S| \leq s} \min_{\Delta \neq 0, \|\Delta_{S^c}\|_1 \leq 3\|\delta_S\|_1} \frac{\|\mathbf{X}\Delta\|_2}{\sqrt{n}\|\Delta_S\|_2} > \kappa(s).$$

This assumption is better understood if we build it step by step. Start by assuming that the Gram or Hessian matrix $\mathbf{X}^\intercal\mathbf{X}/n$ is positive definive, so that the problem admits a unique solution. This assumption is highly

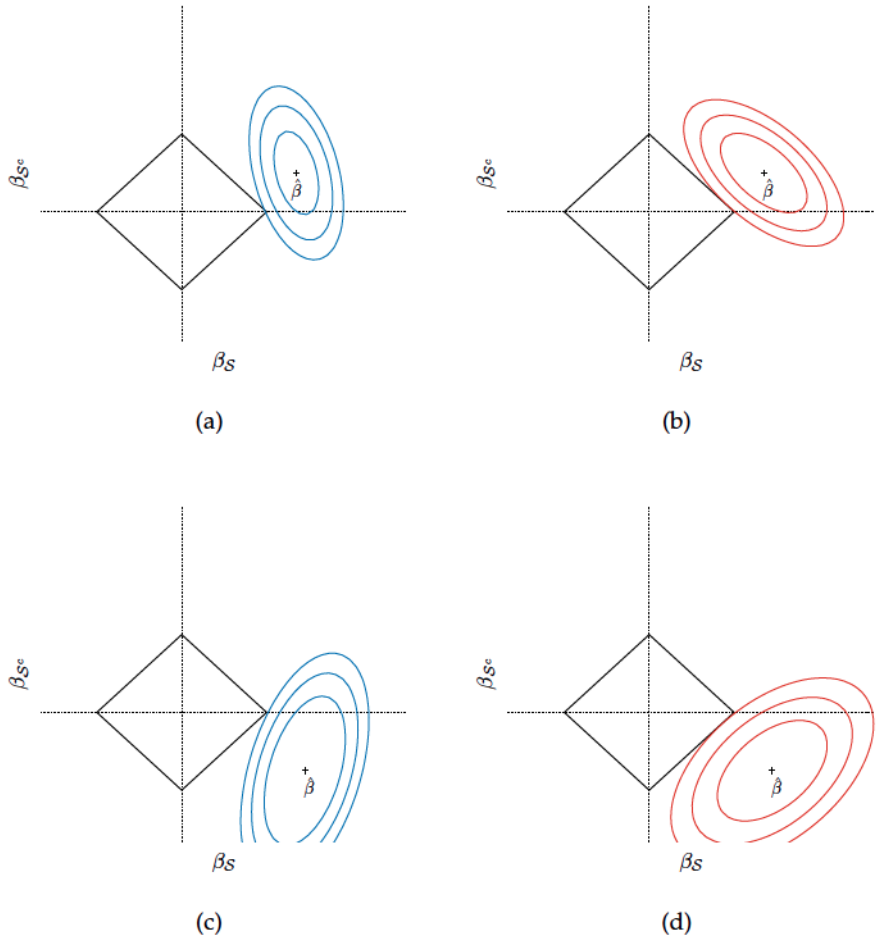


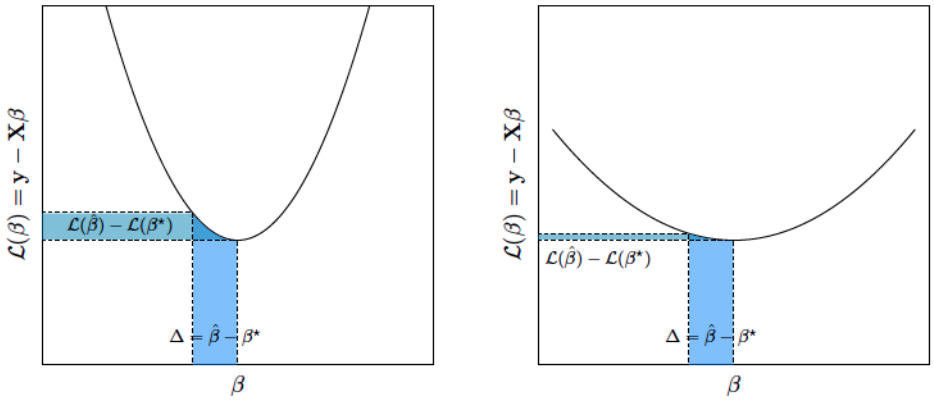

Figure 1.9 – *Configurations (a) and (c) satisfy the irrepresentable condition, configurations (b) and (d) do not.*

unrealistic in high-dimension. If the true signal is sparse, say of sparsity $s$, then the solution is identifiable if and only if all submatrices of size $2s$ of the Hessian are positive definite, that is to say the minimum eigenvalue of all submatrices of size less than $2s$ is positive.

If we are no longer interested in the identification of the true support, but in sharp estimation and prediction properties, then what we need is somewhat stronger than positive eigenvalues, we need large positive eigenvalues. In classical statistical terms and as illustrated by Figure 1.10, we need the Fisher information to be large enough so that an estimation gap $\Delta = \boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}$ induces a difference in likelihood of at least $\kappa\|\Delta\|_2$, or reversely, the smaller the likelihood difference, the smaller the estimation error. In analytical terms, derive a second-order Taylor series expansion near $\boldsymbol{\beta}^\star$ in the direction $\Delta$, to observe that this strong convexity assumption amounts to uniformly lower bound the eigenvalues of the Hessian matrix in the neighborhood of the true parameter $\boldsymbol{\beta}^\star$:

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_n^2 - \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^\star\|_n^2 = -\frac{2}{n}\langle \mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^\star), \Delta\rangle + \|\mathbf{X}\Delta\|_n^2 + o(\frac{\|\Delta\|_2^2}{n})$$

This uniform lower bound is again too strong in high-dimensional settings. Therefore on top of considering reduced size matrices, we focus on




(a) High Fisher information (b) Low Fisher information

Figure 1.10 – *Loss function with high curvature, or Fisher Information, in panel (a), low curvature, or Fisher Information, in panel (b).*

a restricted neighborhood, which is the cone $\{\Delta \in \mathbb{R}^p, \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1\}$, where we know the Lasso error term $\Delta = \boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}$ to reside, hence the denomination *restricted eigenvalue*.

The consequence of the restriction to the cone is that there is no guarantee that the solution will be unique. However, with large probability, all solutions are concentrated within the same $\ell_2$ or $\ell_1$ ball around the true parameter $\boldsymbol{\beta}^\star$. Besides, under supplementary assumptions on the mimimal nonzero value, estimation or prediction bounds can be completed by thresholding steps in order to provide model selection guarantees.

This assumption is the weakest assumption possible, except by a slight modification: change the $\|\Delta_S\|_2$ at the denominator into a $\|\Delta_S\|_1$ to obtain the compatiliby assumption, but we lose the eigenvalue interpretation.

We refer to S. Negahban and Yu (2012) for a generalization of this assumption to adress regularized M-estimators under a larger spectrum of sparsity assumptions on $\boldsymbol{\beta}^\star$.

**How to correctly tune the amount of regularization?**

In the previous sections, the amount of regularization $\lambda_n$ was considered as given, and the question of its choice was purposely eluded. However, when using the Lasso, this is actually the first practical question which arises: what is the correct amount of regularization. All questions related to correct estimation and model selection are actually conditional to the correct choice of $\lambda_n$, since this value roughly speaking determines the size of the model selected by the Lasso. When applying the Lasso, what we obtain is better described via regularization paths: the set of coefficients obtained over varying $\lambda_n$'s, from the null model to the largest possible model given the number of observations available. Most of the time, the Lasso behaves as its Lars (Efron et al. 2004) approximation, adding one variable at a time. However, it sometimes happens that some variables previously added to the set of active variables disappear from the selected model.

Formally, the regularization path provides a collection of models of increasing sizes $\mathcal{M}_\Lambda = \{m_\lambda, \lambda \in \Lambda\}$. As such, $\ell_1$ regularization provides an intelligent way of exploring the much too large set of possible models, which recalls the fact that $\ell_1$ regularization is first and foremost a convex relaxation of the $\ell_0$ regularized problem.

The main issue is to the select the correct amount of regularization and choose a model along the path. Tibshirani (1996) suggests the use of cross-validation. However, the objective of cross-validation is the selection of a model which guarantees to maintain good predictions on new datasets, based on a minimization of an estimation of the generalization error. Yet, cross-validation offer no theoretical guarantees. In comparisaon, the penalized criterion developed in Baraud et al. (2010) addresses the problem of selecting the estimator with smallest Euclidean risk among any family of estimators. In particular, it answers the question of tuning the amount of regularization of Lasso estimators. This criterion is valid under high-dimensional settings and is proved to satisfy non-asymptotic risk bounds under no assumptions on the true model.

Those criteria provide guides on how to tune the amount of regularization in the light of prediction problem [P1] but does not provide a desirable guide in the light of estimation problem [P2] or selection problem [P3]. In chapter 4, though, we happen to need a model with good prediction properties, and resort to the procedure of Baraud et al. (2010).

Thanks to the computation of the Lasso degrees of freedom (Zou et al. 2007, Dossal et al. 2011), BIC and AIC criteria seem adaptable. Yet, their justification rely on asymptotic approximations, which seem highly unrealistic, if not irrelevant, in high-dimensional settings. Extended BIC criteria have been suggested to correct for the Laplace approximation. For a model $S$ of size $s$, denoting by $\widehat{\beta}_S$ the maximum likelihood estimator restricted to model $S$, with corresponding log-likelihood $\ell(\widehat{\beta}_S)$, the extended BIC criterion is defined by

$$EBIC(s) = \ell(\widehat{\beta}_S) - \frac{s}{2}\log n - s \log p.$$

EBIC comes from the addition of a uniform prior on models $S$, such that starting with $p$ variables, each model of size $s$ is given a prior probability of $(p+1)^{-1}(C_p^s)^{-1}$. The consistency of EBIC has been proved in high-dimensional sparse fixed design linear regression (Chen and Chen 2008) and adapted to Gaussian graphical models (Gao et al. 2012, Foygel and Drton 2010).

### 1.3.2 High-dimensional variable selection in GGMs

The inference of a GGM is based upon proposition 1.2, which states that the precision matrix $\Theta$ can in fact be interpreted as the adjacency matrix of an undirected weighted graph $\Gamma$ representing the partial correlation structure between variables $X_1, \ldots, X_p$. Therefore inferring the graph of conditional dependencies $\Gamma$ amounts to recovering the support of $\Theta$ and more than estimating $\Theta$, the main issue in this framework is to answer selection problem [P3] and correctly select the set of nonzero entries of $\Theta$.

### Maximum Likelihood inference

Consider that we observe $n$ identically and independently distributed (hereafter i.i.d.) observations from a multivariate Gaussian distribution with covariance $\Sigma$, which are stored once centered in a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. For each observation $i, (i = 1, \ldots, n)$ and gene $g, (g = 1, \ldots, p)$, entry $x_{ig}$ contains the expression level observed for gene $g$ in the $i$th sample. These $n$ observations must be collected in close enough conditions so that we can assume that they follow the exact same distribution once centered. Independency of the observations also implies that time-course measurements do not fit this undirected model. Chapter 2 is devoted to models designed for time-course, i.e. longitudinal, data.

GGMs fall in the family of exponential models, for which the whole range of classical statistical tools apply. As soon as $n$ is greater than $p$, the model likelihood admits a unique maximum over the set $\mathcal{S}_p^+$, defining a Maximum Likelihood Estimator (MLE). Following the assumption that $X$ is Gaussian, the MLE of $\Theta$ is defined by:

$$\widehat{\Theta}^{\text{MLE}} = \underset{\Theta \in \mathcal{S}_p^+}{\arg\max} \quad (2\pi)^{-p/2} \det(\Theta) \exp\left(-\frac{1}{2}\mathbf{X}^\intercal \Theta \mathbf{X}\right). \tag{1.9}$$

Let us denote the empirical covariance matrix by $\widehat{S} = \mathbf{X}^\intercal\mathbf{X}/n$. After log-transformation and use of the Trace operator property $\text{Tr}(a^\intercal b) = \text{Tr}(ba^\intercal)$, for every compatible vectors $a$ and $b$, Problem 1.9 becomes:

$$\widehat{\Theta}^{\text{MLE}} = \underset{\Theta \in \mathcal{S}_p^+}{\arg\max} \quad \log\det(\Theta) - \langle \Theta, \widehat{S} \rangle, \tag{1.10}$$

where $\langle A, B \rangle$ denotes the matrix inner-product associated with the Frobinius norm, $\langle A, B \rangle = \text{Tr}(A^\intercal B)$.

When $n$ is larger than $p$, Problem 1.10 admits a unique solution equal to $\widehat{S}^{-1}$. As the square product of a centered and scaled Gaussian vector, $\widehat{S}$ follows a Wishart distribution, which is the multivariate generalization of the chi-square distribution. As a result, its inverse $\widehat{S}^{-1}$ naturally follows an inverted Wishart distribution whose parameters admit an analytical close form.

There are two major limitations with the MLE regarding the objective of graph reconstruction by recovering zeroes in the estimate of $\Theta$. First of all, and not of little importance, we need $n$ to be larger than $p$ to be able to even define this estimator, which is never the case in microarray studies, unless we focus on a very restrained subset of candidate genes. Second, even in the case where we would be lucky enough to gather enough data, the MLE provides an estimate of the saturated graph: all genes are connected to each other, which is of no interest at all.

What saves us here is a common property of biological networks, namely sparsity: among all $p(p-1)/2$ possible interactions between genes, only a few actually take place. Sparsity makes the estimation feasible in the case where $n$ is smaller than $p$ since we can concentrate on sparse or shrinkage estimators with less degrees of freedom than in the original problem. Henceforth, the question of selecting the correct set of edges in the graph is treated as a question of model (or covariate) selection.

### Background on High-Dimensional Inference of GGM

The different methods for model selection/estimation in GGMs roughly fall into three categories. The first contains constraint-based methods, performing statistical tests. We mention that the procedure in Drton and Perlman (2007; 2008) relies on asymptotic considerations, a regime never attained in real situations. The forward selection method combined with permutation tests suggested in Kiiveri (2011) would fall into this category. Limited-order partial correlations were also considered in Wille and Bühlmann (2006), Castelo and Roverato (2006). The second of these categories is composed of Bayesian approaches, see for instance Dobra et al. (2004), Jones et al. (2005), Rau et al. (2011). However, constructing priors on the set of concentration matrices is not a trivial task and the use of MCMC procedures limits the range of applications to moderate-sized networks. The third category contains regularized estimators, which add a penalty term to the likelihood in order to reduce the complexity or degrees of freedom of the estimator. A first shrinkage estimator was proposed by Schäfer and Strimmer (2005). This approach consists in using a weighted average of two different estimators, the first being unconstrained (thus having small bias but large variance), the second being low-dimensional (and thus exhibiting small variance but large bias).

Let us now introduce adaptations of $\ell_1$-regularized procedures to the inference of high-dimensional GGMs. In Meinshausen and Bühlmann (2006), a first attempt was made under the name of *neighborhood selection*. This approach solves $p$ different Lasso regression problems, where $p$ is the number of genes in the network. Subsequently two other articles, Banerjee et al. (2008) and Yuan and Lin (2007a), independently provided an improvement of the initial work of Meinshausen and Bühlmann (2006). In both works, the problem is seen as a penalized maximum likelihood (PML) problem and is solved as a recursive 'Lasso-like' problem. The next improvement in this vein comes with the *Graphical* Lasso, or `gLasso`, of Friedman et al. (2008), which makes this penalized likelihood approach highly attractive in terms of computational cost, with very recent improvements for high-dimension developed in Mazumder and Hastie (2011). Still, the neighborhood selection approach remains a lot cheaper, computationally speaking.

### Highlights on $\ell_1$ Regularizers for GGMs

Let us review in little more details the two Lasso-type techniques on which we build upon in this thesis, namely the *neighborhood selection* and the *Graphical-Lasso* approaches.

On the one hand, the $\ell_1$-penalized estimator, proposed in Banerjee et al. (2008) and advantageously solved by the `gLasso` algorithm, directly considers the original penalized likelihood problem:

$$\widehat{\Theta}^{\lambda} = \arg\max_{\Theta \in \mathcal{S}_p^+} \log\det(\Theta) - \langle \Theta, \widehat{S} \rangle - \lambda \, \|\Theta\|_{\ell_1}. \tag{1.11}$$

In this regularized problem, the $\ell_1$-norm on the entries of the concentration matrix drives some coefficients to zero: it enforces sparsity. The non-negative parameter $\lambda$ tunes the global amount of sparsity: the larger the

parameter $\lambda$, the fewer edges in the graph. A large enough penalty level produces an empty graph. As $\lambda$ decreases towards zero, the estimated graph tends towards the saturated graph and the estimated concentration matrix tends towards the usual unpenalized MLE $\widehat{\Theta}^{\text{MLE}}$. By construction, this approach guarantees a well-behaved estimator of the concentration matrix, that is to say sparse, symmetric and positive-definite.

On the other hand, the more naive neighborhood selection procedure has been reported to be more accurate in terms of edge detection. The reader is referred to Villers et al. (2008) and Rocha et al. (2008). This approach determines the graph of conditional dependencies $\Gamma$ by solving a series of $p$ independent $\ell_1$-penalized regression problems, successively estimating each gene neighborhood. Recall that $\mathbf{X}$ is the $n \times p$ matrix of observations, with column $g$ containing the vector $\mathbf{X}_g$ of $n$ observations for gene $g$. Matrix $\mathbf{X}_{\setminus g}$ contains all columns $\mathbf{X}$ except its $g$th column, that is to say observations on all genes except expression levels of gene $g$. Concretely, for each gene $g$, expression levels are "explained" by the expression levels of remaining genes. Neighbors of gene $g$ in the graph $\Gamma$ are estimated by the nonzero elements of $\hat{\beta}_g$ solving Problem 1.12.

$$\hat{\beta}_g = \underset{\beta \in \mathbb{R}^{p-1}}{\arg\min} \left\| \mathbf{X}_g - \mathbf{X}_{\setminus g}\beta \right\|_n^2 + \lambda \, \|\beta\|_1. \tag{1.12}$$

Indeed, if $\text{ne}(g)$ denotes the set of neighbors of gene $g$ in the graph of conditional dependencies $\Gamma$ associated to the concentration matrix $\Theta$, then proposition 1.1 implies that the best linear approximation of the random vector $X_g$ by remaining gene expressions $X_{\setminus g}$ is given by:

$$X_g = \sum_{h \in \text{ne(g)}} \beta_{gh} X_h = - \sum_{h \in \text{ne(g)}} \frac{\Theta_{gh}}{\Theta_{gg}} X_h.$$

As a result, Problem 1.12 aims to estimate coefficients $\beta_{gh}$ proportional to the concentration matrix entries of interest $\Theta_{gh}$.

Actually, solving the $p$ regression problems defined by (1.12) may be interpreted as inferring the concentration matrix in a penalized maximum *pseudo*-likelihood framework, as depicted in Rocha et al. (2008), Ambroise et al. (2009), Ravikumar et al. (2010): the joint distribution of $X$ is approximated by the product of the $p$ distributions of the $p$ variables, conditional on the other ones, as if these distributions were independent, that is

$$\mathcal{L}(\Theta; \mathbf{X}) = \sum_{g=1}^{p} \sum_{i=1}^{n} \log \mathbb{P}(x_{ig} | \mathbf{X}_{i\setminus g}; \Theta_g),$$

where $\mathbf{X}_{i\setminus g}$ is the $i$th observation of the vector $X$ deprived of the $g$th coordinate. This pseudo-likelihood is based upon the (false) assumption that conditional distributions of expression levels are independent. Particularly the distribution of gene $g$ expression levels conditional on gene $h$ is assumed independent from the distribution of gene $h$ conditional on gene $g$, ignoring the symmetry condition on concentration matrices. Because the neighborhoods of the $p$ genes are selected separately, a post symmetrization must be applied to manage inconsistencies between edge selections; Meinshausen and Bühlmann (2006) suggests AND or OR rules.

As for the comparison of theoretical properties, even though there is no reason for the Graphical Lasso and the neighborhood selection approach to result in identical estimates at fixed $n$, they are both shown asymptotically consistent in terms of edge detection (as $n$ goes to infinity) under their respective strong but necessary irrepresentability assumptions. Ravikumar et al. (2011) provide an irrepresentable or mutual incoherence assumption similar to 1.8 for the graphical Lasso. The Hessian $\mathbf{X}^\intercal\mathbf{X}/n$ of the least square problem is naturally replaced by the Hessian of Problem 1.11, namely $H = \Sigma \otimes \Sigma$. This Hessian corresponds in fact to a covariance matrix at the edge level, since for every pair of edges $(i,j),(k,\ell)$, $H_{(i,j),(k,\ell)} = \mathrm{cov}(X_iX_j, X_kX_\ell)$. Therefore, as irrepresentable condition for linear regression involved covariances between relevant and irrelevant covariates, the irrepresentable condition for the graphical Lasso involves the covariances between relevant and irrelevant edges. Denoting by $\mathcal{E}$ the set of true edges $\mathcal{E} = \{(g,h) \in \mathcal{V}^2, g \neq h \text{ and } \Theta_{gh} \neq 0\}$, we can quote the corresponding irrepresentable condition.

**Definition 1.11** (Irrepresentable condition for the graphical Lasso Ravikumar et al. (2011)) *There exists $\eta > 0$ such that*

$$\max_{e \in \mathcal{E}^c} \|H_{e\mathcal{E}} H_{\mathcal{E}\mathcal{E}}^{-1}\|_1 \leq 1 - \eta$$

The conclusion of Ravikumar et al. (2011) is that the main differences in terms of performances between the two methods from an information theoretic point of view lies in this hypothesis. However, despite two particular examples where the irrepresentable condition for neighborhood selection is seen less restrictive than its graphical Lasso counterpart, there is no general rule to be known.

## 1.4 A Stepping-Stone Towards the Structured Modeling and Inference of High-Dimensional GGMs

The introduction of $\ell_1$ regularization has rendered possible the address of high-dimensional problems, which are out of reach of classical asymptotic theory based upon the paradigm of small $p$, large $n$. However, $\ell_1$ regularization reaches its limits in terms of robustness: support recovery is only guaranteed under strong assumptions without which a large number of false positives can be introduced. In particular, $\ell_1$ regularization suffers from designs with correlated covariates. This observation has motivated the development of a vivid field of research building variations on $\ell_1$ regularization in three main directions: combinations of $\ell_1$ regularization and bootstrap sampling, weighted $\ell_1$ regularizers and a large range of sparsity-inducing norms. While the notion of *sparsity* triggered the introduction of $\ell_1$ regularization, most of the variations on $\ell_1$ regularization are stimulated by the idea that the sparsity pattern follows a particular *structure*. By resorting to sparsity, $\ell_1$ regularization makes it possible to answer statistical problems that were originally unthought of. By instilling structure, variations on $\ell_1$ aim at perfectly fitting the underlying structure of the data, and thereby increase the robustness of the answer.

**Bootstrap Methods.** The first statistical answer to the lack of robustness presented by the Lasso comes from bootstrap sampling, in the spirit of the Bolasso (Bach 2008a), or stability selection (Meinshausen and Bühlmann 2010) in the field of high-dimensional GGM. The idea underlying the Bolasso is that for a well-chosen amount of sparsity, the Lasso is most likely to select all true covariates with a probability tending to one, while most unrelevant covariates are selected with a non zero but strictly less than one probability. If we infer the model on various bootstrap samples, each of the estimated support will include the true model with probability tending to one, along side a few false positives. Therefore, taking the intersection of those supports should discriminate the true support from false inclusions. Stability selection takes a softer stand, by defining for each covariate the probability that it is selected across the set of bootstrap samples for a given amount of regularization, and retains covariates reaching a certain selection probability. Even though it requires some computing time, these bootstrap corrections have been recognized to improve the accuracy of selected models (Haury et al. 2011b, Rohart 2011).

**Weighted $\ell_1$ Regularization.** Another approach developed to reduce the inclusion of false positives is the fine tuning of the amount of regularization, covariate by covariate. The adaptive Lasso (Zou 2006, Zhou et al. 2011), which corrects the penalty level $\lambda_n$ by weights inversely proportional to an initial estimator $\hat{\boldsymbol{\beta}}_{\text{init}}$, aims at reducing the bias on large coefficients while reducing the probability of falsely selecting unrelevant covariates. Instead of adapting the weights according to an initial estimate, Ambroise et al. (2009) suggest to adapt weights according to the specific biological structure of the data. Focusing on the inference of high-dimensional GGMs from i.i.d. transcriptomic data, their idea is to modulate the penalty levels according to the topological structure of the network. Chapter 2 adapts this idea to the inference of GGMs from time-course transcriptomic data.

**Sparsity-inducing Regularizations.** Last but not least, we must mention that a wide variety of sparsity-inducing norms have already designed in order to tackle as many different statistical issues as there are of concrete application frameworks. In front of this outburst of new regularizers, references S. Negahban and Yu (2012), Bach (2010), Bach et al. (2012) form attempts at the definition of a generalized theory, to replace the case by case analysis of each new suggestion. Some of them require our attention. First of all, the elastic-net (Zou and Hastie 2005) provides an answer to the problematic case of correlated subsets of covariates. By combining the $\ell_1$ norm to an $\ell_2$ norm, the elastic-net aims to select all correlated covariates as one, contrary to the Lasso, which would let those covariates compete to enter the model. As a result, the elastic-net has been regarded as a good solution to stabilize the support recovered by the Lasso in the case of correlated designs (Allassonnière and Giraud 2011). Beyond sparsity requirements, the modeling of redondancies through low rank matrices has been implemented through nuclear norm regularizers, allowing the decomposition of matrices into a low rank and a sparse components. Finally, group-sparse regularizers have been designed to tackle what is

known in machine learning as multi-task settings, or could correspond to panel datasets: redondant datasets, called *tasks*, are collected about the same phenonemon, be it multiple cameras, multiple sensing channels, multiple individuals, correlated covariates. The objective is to combine those redundant tasks under the hypothesis that they share the same signal sparsity pattern, without merging them into a single dataset as if they were strictly i.i.d. Chapter 3 focuses on this particular question.

# 2 Weighted-Lasso for Structured Network Inference from Time Course Data

We present a weighted-Lasso method to infer the parameters of a first-order vector auto-regressive model that describes time course expression data generated by directed gene-to-gene regulation networks. These networks are assumed to own a topological structure which helps define a weighted $\ell_1$ regularization. This prior structure can be either derived from expert biological knowledge or inferred by the method itself. We illustrate the performance of this structure-based penalization both on synthetic data and on two canonical regulatory networks (the yeast cell cycle regulation network and the E. coli S.O.S. DNA repair network).

This chapter is mainly inspired from reference Charbonnier et al. (2010).

Contents

## 2.1 Introduction

Many transcriptomic datasets do not fit the i.i.d. settings at all, notably time course expression datasets. Assuming a first-order vector autoregressive ($VAR1$) model, several authors have already provided inference methods handling high-dimensional settings: Opgen-Rhein and Strimmer (2007) suggested a shrinkage estimate while Lèbre (2009) performed statistical tests on limited-order partial correlations to select significant edges. In a recent work, Shimamura et al. (2009) proposed to deal with this $VAR1$ setup by combining ideas from two major developments of the Lasso to define the Recursive elastic-net. As an elastic-net (Zou and Hastie 2005), this method adds an $\ell_2$ penalty to the original $\ell_1$ regularization, thus encouraging the simultaneous selection of highly correlated covariates on top of the automatic selection process due to the $\ell_1$ norm. As in the adaptive-Lasso (Zou 2006), weights are corrected on the basis of a former estimate so as to adapt the regularization parameter to the relative importance of coefficients. Note that, in this context, we are no longer looking for an estimate of the inverse of the covariance matrix but of the parameters of the $VAR1$ model, which leads to a *directed graph*.

In this chapter, we aim to couple $VAR1$ modeling of time course data with an $\ell_1$-regularized approach taking the topological structure of the network into account. A simple example of topological structure would split the genes into two groups: a group of *hubs* that exhibit a high connection probability to all other genes and a group of *leaves* that only receive edges leaving from the hub class. This information can either be inferred or recovered from biological expertise since recovering hubs consists roughly in exhibiting *transcription factors* in regulatory networks, a large number of them being already identified by the biologists.

Another refinement of our method is to built on the adaptive-Lasso (Zou 2006, Zhou et al. 2009) which is known to reduce false positive rate compared to the classical Lasso. As such, our method belongs to the larger family of weighted-Lasso methods. Shimamura et al. (2007) build upon Meinshausen and Bühlmann (2006)'s neighborhood selection and the adaptive-Lasso to improve inference of networks in an i.i.d. context. They choose separate penalties for each neighborhood selection problem and adapt each individual penalty coefficient to the information brought by an initial ridge estimate. Here, we suggest to lower the bias of the Lasso by not only using information from an initial statistical inference but also from prior knowledge about the topology of the network that assumes the existence of genes with high connection probability to other genes.

The rest of the chapter is organized as follows: in the next section, the $VAR1$ model and associated likelihood function are briefly recalled; an $\ell_1$-penalized criterion is proposed where each parameter of the $VAR1$ model, representing the graph of interest, is weighted according to a prior structure of the network. The weights can also depend on a previous estimate just as in the adaptive-Lasso. In Section 3, the inference procedure is detailed: we present how the topological structure can be recovered; from that point on, network inference reduces to a convex optimization problem which we solve through an active-set algorithm based upon the approach of Osborne et al. (2000). Finally, an experimental Section inves-

tigates the performances of the method. First, simulated data are considered; then, we try to recover edges implied in two different regulation processes. First in yeast cell cycle, by analyzing the Spellman et *al.*'s dataset and comparing the selected edges to the direct regulations collected from the `Yeastract` database; second in *E. coli*, by analyzing U. Alon's precise kinetic data on S.O.S. DNA repair subnetwork.

## 2.2 Modeling Structured Regulation Networks from Time-Course Data

### 2.2.1 Auto-Regressive Model and Sparse Networks

The dynamics of RNA measurements $X_0, X_1, \ldots, X_T$ of $p$ genes at $T+1$ regular time points are represented by a first-order vector autoregressive model $VAR1$ as in equation (2.1). Each measurement $X_t$ is a size $p$ row vector containing the expression levels of the $p$ genes of interest at time $t$.

$$X_t = X_{t-1}A + \varepsilon_t, \quad \text{for all } t \geq 1, \tag{2.1}$$

where matrix $A = (A_{gh})_{g,h\in\mathcal{P}}$ is a $p \times p$ matrix governing the dynamics of expression levels over time. To guarantee that $A$ is stationnary and in its canonical representation we assume that $A$ has eigenvalues of absolute values strictly smaller than 1. Variations away from these dynamics are captured by a white Gaussian noise $\{\varepsilon_t\}_{t=1,\ldots,T}$ satisfying for every $t, s \geq 1$ assumptions (2.2) and (2.3).

$$\mathbf{E}(\varepsilon_t) = 0, \tag{2.2}$$

$$\mathbf{E}(\varepsilon_t^\intercal \varepsilon_s) = \delta_{st}\sigma^2 I_p. \tag{2.3}$$

Under these assumptions, $\{X_t\}_{t=0,\ldots,T}$ follows a first-order Markov process homogeneous in time: if expression levels vary over time according to equation (2.1), the regulatory structure among these expression levels is assumed constant over time, as illustrated by Figure 2.1.

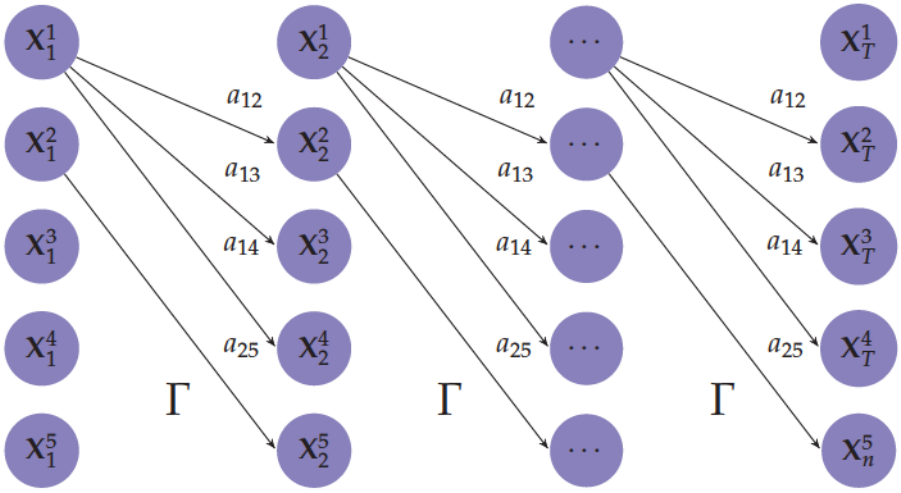



Figure 2.1 – *Example of homogeneous Markov process on a set of five genes.*

**Implications in terms of data collection and normalization** First, because of the homogeneity assumption, $VAR1$ models apply to dynamic measurements but do not provide dynamic networks. Regulations are assumed to be constant over time. Therefore, this model is better suited to draw a picture of short-term regulation dynamics based upon measurements taken at close time points to guarantee the detection of dependencies between time-points and over a short period of time to satisfy the homogeneity assumption. Models taking into account possible evolutions of the regulatory networks over time and better suited for life cycle datasets were for instance developed in Lèbre et al. (2010). Second, assumption (2.3), stating that there is no correlation between contemporaneous noise terms, is only reasonable if no important gene, particularly any gene regulating multiple genes in the dataset, has been omitted and if the data have been correctly normalized, thereby annihilating any correlated measurement errors over the microarrays.

**Network modeling** In this setting, matrix $A$ plays the role of the concentration matrix $\Theta$ in the i.i.d. framework presented in the previous chapter. Indeed, each entry $a_{gh}$ is proportional to the partial correlation coefficient between variables $X_t^g$ and $X_{t-1}^h$, that is to say between the expression of gene $g$ at time $t$ and the expression of gene $h$ at the previous time point, conditional on all other gene expressions at time $t-1$, as expressed in equation (2.4).

$$A_{gh} = \frac{\text{cov}\left(X_t^g, X_{t-1}^h | X_{t-1}^{\setminus h}\right)}{\text{var}\left(X_{t-1}^h | X_{t-1}^{\setminus h}\right)} \propto \frac{\text{cov}\left(X_t^g, X_{t-1}^h | X_{t-1}^{\setminus h}\right)}{\sqrt{\text{var}\left(X_t^g | X_{t-1}^{\setminus h}\right) \text{var}\left(X_{t-1}^h | X_{t-1}^{\setminus h}\right)}} \tag{2.4}$$

Note that Assumption 2.3 bears the important consequence that conditional on the past, contemporaneous gene expressions are necessarily independent.

**Proposition 2.1** (Absence of contemporaneous partial correlations) *Assume* (2.1), (2.2) *and* (2.3). *For every pair of genes* $(g,h)$ *and time point* $t$*:*

$$cov(X_t^g, X_t^h | X_{t-1}) = 0.$$

As in the i.i.d. setting, nonzero entries of $A$ code for a graph describing the conditional dependencies between gene expression levels, except that the graph is now directed, as in Figure 2.2. Even though time is omitted, the graphical representation in Figure 2.2 is in fact equivalent to the homogeneous Markovian representation of Figure 2.1. Proposition 2.1 clearly states that no regulation can exist between contemporaneous points. An edge from $h$ to $g$ is added to the graph if, conditional on all gene expressions except gene $h$ at time $t-1$, the covariance between $X_{g,t}$ and $X_{h,t-1}$ is nonzero. Identifying the nonzero entries of $A$ is again equivalent to reconstructing the graph of conditional dependencies. However, there are two main differences between this dynamic version of partial correlation and the notion of partial correlation expressed in the previous chapter. First, the conditioning is made upon all gene expressions from

the previous time-point, therefore self-loops are allowed. Second, the correlation considered between the two genes is asymmetric: we consider the correlation between the past expression levels of gene $h$ and the present expression levels of gene $g$, leading naturally to an asymetric matrix of partial correlations and a directed graph of conditional dependencies.

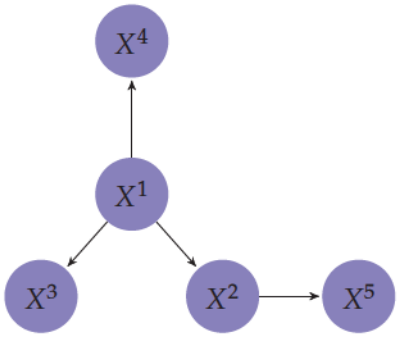



Figure 2.2 – *Graph of conditional dependencies associated with the homogenous Markov process represented in Figure 2.1.*

**Estimation in the case** $p < T$ Denote by $\mathbf{X}$ the $(T+1) \times p$ matrix of centered, scaled to unit-variance data, whose $t$th row contains the information $X_t$ relative to the $p$ variables at time $t$. The empirical variance-covariance matrix $\mathbf{S}$ and the empirical temporal covariance matrix $\mathbf{V}$ are then given by

$$\mathbf{S} = \frac{1}{T}\mathbf{X}_{\setminus T}^\intercal \mathbf{X}_{\setminus T}, \quad \mathbf{V} = \frac{1}{T}\mathbf{X}_{\setminus T}^\intercal \mathbf{X}_{\setminus 0},$$

where $\mathbf{X}_{\setminus k}$ denotes matrix $\mathbf{X}$ deprived of its $k$th row.

Thanks to the assumptions we make on the modeling, the log-likelihood of the $VAR(1)$ factorizes into a simple expression:

$$\begin{aligned}
\mathcal{L}(A; \mathbf{S}, \mathbf{V}) &= \sum_{t=1}^{T} \log f(X_t | X_{t-1}) \\
&= -\frac{1}{2\sigma^2} \sum_{t=1}^{T} (X_t - AX_{t-1})^\intercal (X_t - AX_{t-1}) + c \\
&= -\frac{1}{2\sigma^2} \sum_{t=1}^{T} X_{t-1}^\intercal A^\intercal A X_{t-1} - 2X_t^\intercal A X_{t-1} + c \\
&= -\frac{T}{2\sigma^2} \left[ \mathrm{Tr}\left(A^\intercal \mathbf{S} A\right) - 2\mathrm{Tr}\left(\mathbf{V}^\intercal A\right) \right] + c,
\end{aligned}$$

As a result, the maximum likelihood estimator (MLE) of $A$ is easily recovered and recalled in the following proposition.

**Proposition 2.2** (Maximum Likelihood Estimator) *Assume $p < T$. Then $\mathbf{S}$ is invertible and maximizing the log-likelihood of the VAR1 process is equivalent to the following maximization problem*

$$\max_{A \in \mathcal{M}_p(\mathbb{R})} \left\{ \mathrm{Tr}\left(\mathbf{V}^\intercal A\right) - \frac{1}{2}\mathrm{Tr}\left(A^\intercal \mathbf{S} A\right) \right\},$$

*whose solution is given by*

$$\widehat{A}^{\mathrm{mle}} = \mathbf{S}^{-1}\mathbf{V}. \tag{2.5}$$

**Remark 2.1** *Thanks to the assumptions we made on noise terms the VAR1 model can be factorized and seen as a usual regression problem. Denote by $\mathbf{X}_p$ (respectively $\mathbf{X}_f$) the $T$ first (respectively last) rows of $\mathbf{X}$. $\hat{A}^{ols}$ is naturally given by $(\mathbf{X}_p^\intercal \mathbf{X}_p)^{-1}\mathbf{X}_p^\intercal \mathbf{X}_f = \mathbf{S}^{-1}\mathbf{V} = \hat{A}^{mle}$. The MLE (2.5) is straightforwardly equivalent to the ordinary least square estimate (OLS) of $A$.*

**Estimation in the high-dimensional case** Solution (2.5) requires a covariance matrix $\mathbf{S}$ that is invertible, which occurs when $\mathbf{S}$ is at least of rank $p$. In real situations the actual number of observations $T$ is often about or lower than the number of variables, thus the MLE needs to be regularized. Regularization such as Moore-Penrose pseudo inversion or $\ell_1$-regularization can be applied on matrix $\mathbf{S}$ in order to make the inversion always achievable. A sharpest approach is investigated in Opgen-Rhein and Strimmer (2007), where the OLS solution is regularized by shrinking both matrices $\mathbf{S}$ and $\mathbf{V}$.

We suggest to draw inspiration from the $\ell_1$-penalized likelihood approach developed by Banerjee et al. (2008) in the case of i.i.d. samples of a multivariate Gaussian distribution: here, samples are no longer i.i.d yet linked through time by the *VAR*1 model. Still, the sparsity can be controlled with a positive scalar $\lambda$ adjoined to an $\ell_1$-norm penalty on $A$ by solving

$$\hat{A}^{\ell_1} = \arg\max_A \left\{ \mathrm{Tr}\left(\mathbf{V}^\intercal A\right) - \frac{1}{2}\mathrm{Tr}\left(A^\intercal \mathbf{S} A\right) - \lambda \|A\|_1 \right\}, \qquad (2.6)$$

where the $\ell_1$-norm of matrix $A$ is simply defined by $\|A\|_1 = \sum_{i=1}^p \sum_{j=1}^p |A_{ij}|$ . Since MLE and OLS are equivalent in this framework, solution to the penalized-likelihood formulation (2.6) is equivalent to solving $p$ independent Lasso problems on each column of $A$, which is exactly Meinshausen and Bühlmann (2006)'s approach. The difference is that it does not require any post-symmetrization since there is no symmetry constraint on $A$ in the present context.

### 2.2.2 A Structured Modeling of the Network

As in reference Ambroise et al. (2009), we suggest that the graphical representation of $A$ owns a particular topological structure which identifies clusters of genes with characteristic connectivity patterns. Indeed, the $\ell_1$-norm regularization encourages a first restriction on the network's topology inferred through criteria (2.6), by encouraging sparsity. Yet, it is well known that, by penalizing truly significant entries of $A$ as much as truly zero entries, a single $\ell_1$ penalization leads to biased estimates and a particularly strong number of false positives (Knight and Fu 2000, Zou 2006). Weighted-Lasso approaches can lower this bias by adapting penalties to prior information about where the true zero entries should be, relying on possibly data-driven as well as biological information. An existing correction is given by the adaptive-Lasso (Zou 2006, Zhou et al. 2009). Penalty coefficients are alleviated or increased using individual weights reversely proportional to a consistent initial estimate $A^{\text{init}}$.

The main purpose of this chapter is to show the interest of taking into account information about the topology of the network: not only should

we scale coefficients individually, but also consider the underlying organization of the gene set $\mathcal{P}$. Adaptation of weights is made by providing $A$ with a well-chosen prior distribution, relying on the organization of $\mathcal{P}$. We assume that genes are spread through a partition of $\mathcal{P}$ into $\mathcal{Q}$ classes of connectivity. Both existences and weights of edges, described by the elements of $A$, depend on the connectivity class each vertex belongs to. Denote by $Z_{iq}$ the indicator function that gene $i$ belongs to class $q$. Conditional on the fact that gene $i$ belongs to cluster $q$ and gene $j$ belongs to cluster $\ell$, each entry $A_{ij}$ is provided with an independent prior distribution $f_{ijq\ell}$. Following Ambroise et al. (2009), we choose Laplace distributions for $f_{ijq\ell}$ since it is the corresponding log-prior distribution to the $\ell_1$ term in the Lasso. Hence, by choosing

$$f_{ijq\ell}(A_{ij}) = \frac{1}{2\rho_{ijq\ell}} \exp\left\{-\frac{|A_{ij}|}{\rho_{ijq\ell}}\right\},$$

where $\rho_{ijq\ell}$ are scaling parameters, we expect a model whose log-likelihood will naturally make a specific $\ell_1$-penalization term appear.

The interpretation of $\ell_1$ regularizations as Laplace Bayesian prior distribution has been discussed in Gribonval (2011). Despite the fact that many other prior distributions could lead to the same MAP expression, the Laplacian interpretation is rather intuitive. As illustrated on Figure 2.3, the larger the regularization weight $\rho_{ijq\ell}$, the stronger the concentration of the prior probability around 0.

**Modeling star-shaped (or hub) networks.** Many configurations fit into this general model. Ambroise et al. (2009) focused on an affiliation model. This structure opposes intra to inter-cluster connections, assuming the former to be far more likely than the latter. In the present context, where dynamic regulatory networks are represented by directed graphs, the affiliation model unnaturally assumes symmetric probabilities for "incoming" and "outgoing" edges and should be banished. Indeed, adjacency matrices associated to directed gene regulatory networks are asymmetrical. A typical structure consists of star-shaped networks, in which genes belong to two completely different groups. While a group of hubs exhibits a high connection probability to all other genes, the remaining set of genes almost only receives edges leaving from the first class. Illustration of this phenomenon by a gene regulatory network reconstructed on the basis of biological experimentations and computational biology techniques in the budding yeast is presented in Section 2.4. This setup can be summarized as follows:

$$f_{ijq\ell} = \begin{cases} f_{\text{hub}}\left(\cdot;\rho_{\text{hub}}\right) & \text{if } q \text{ is the hub class,} \\ f_{\text{leaf}}\left(\cdot;\rho_{\text{leaf}}\right) & \text{if } q \text{ is not the hub class.} \end{cases}$$

Note that this structure only differentiates edges on the basis of their origin, whether they leave from a hub or not, whatever be the cluster of their arrival points. In this type of structure built around hubs, the number of clusters is fixed at 2.

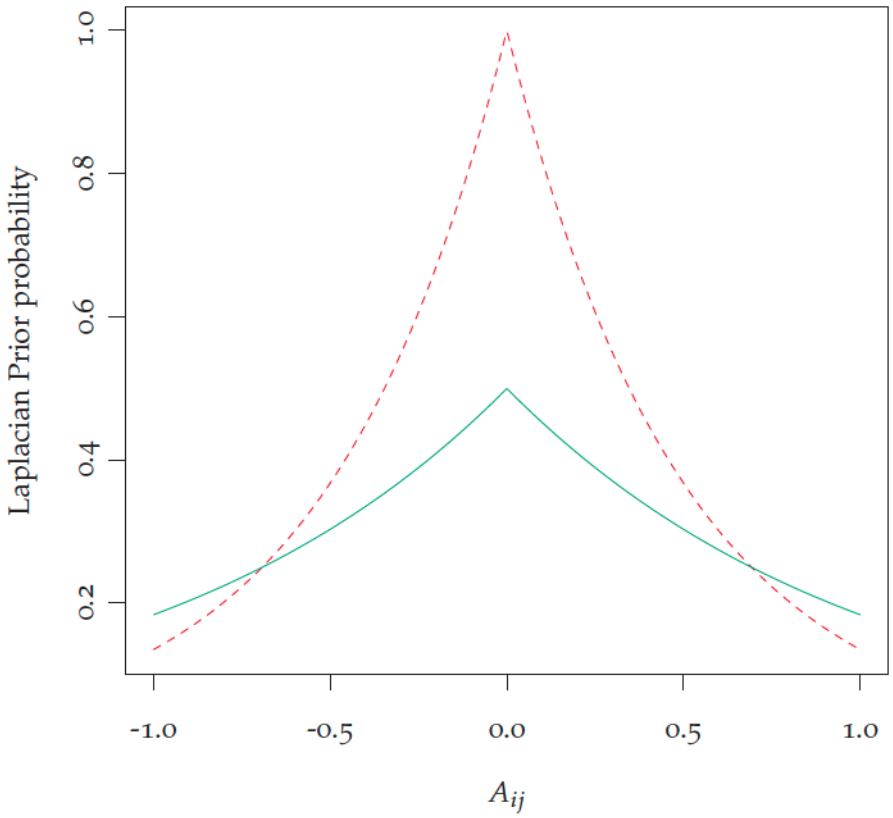



Figure 2.3 – *Laplace distributions for two parameters $\rho_{ijq\ell}$. The larger value leads to the distribution in dashed red line, highly peaked at 0, while the smaller value leads to the distribution in plain green line, more even spread over non zero values.*

Allowing for individual prior information about $i$ and $j$, this model can be generalized to

$$f_{ijq\ell} = \begin{cases} f_{\text{hub}}\left(\cdot; \rho_{ij}\rho_{\text{hub}}\right) & \text{if } q \text{ is the hub class,} \\ f_{\text{leaf}}\left(\cdot; \rho_{ij}\rho_{\text{leaf}}\right) & \text{if } q \text{ is not the hub class.} \end{cases}$$

**The Likelihood.** As the matrix $A$ has been given a prior distribution, our aim is to maximize the posterior probability of $A$, given the data $\mathbf{X}$. For a fixed structure $\mathbf{Z}$, this is equivalent to maximizing the joint probability

$$\hat{A} = \arg\max_{A} \log \mathbb{P}(\mathbf{X}, A; \mathbf{Z}).$$

Now, the likelihood $\mathbb{P}(\mathbf{X}, A; \mathbf{Z})$ is straightforwardly given by

$$\log \mathbb{P}(\mathbf{X}, A; \mathbf{Z}) = \text{Tr}\left(\mathbf{V}^\intercal A\right) - \frac{1}{2}\text{Tr}\left(A^\intercal \mathbf{S} A\right) - \|\mathbf{P}^{\mathbf{Z}} \star A\|_1 + c, \tag{2.7}$$

where $c$ is a constant term and the $p \times p$ penalty matrix is defined by

$$\mathbf{P}^{\mathbf{Z}} = (P^{\mathbf{Z}}_{ij})_{i,j\in\mathcal{P}} = \sum_{q,\ell\in\mathcal{Q}} \frac{Z_{iq}Z_{j\ell}}{\rho_{ijq\ell}}. \tag{2.8}$$

Assuming a star-shaped structure, Equation (2.8) takes the form of Equation (2.9)

$$\mathbf{P}^{\mathbf{Z}}_{ij} = \rho^{-1}_{ij} \cdot \left(\rho^{-1}_{\text{hub}} Z_{i,\text{hub}} + \rho^{-1}_{\text{leaf}} Z_{i,\text{leaf}}\right) = \lambda \cdot \lambda_{ij} \cdot \left(\lambda_{\text{hub/leaf}} Z_{i,\text{hub}} + Z_{i,\text{leaf}}\right), \tag{2.9}$$

where $\lambda > 0$ is a common factor to $\rho^{-1}_{\text{hub}}$ and $\rho^{-1}_{\text{leaf}}$, which can vary so as to adapt overall sparsity of the network while the ratio $\rho^{-1}_{\text{hub}}/\rho^{-1}_{\text{leaf}} = \lambda_{\text{hub/leaf}} < 1$ governs the deflation of the penalty on edges leaving from hubs. Coefficient $\lambda_{ij}$ can be held fixed at 1 when no individual information is taken into account or replaced by any well-chosen transformation of an initial estimate of $A$ in order to provide accurate information on where true zeros might be.

## 2.3 Inference Strategy

Reference Ambroise et al. (2009) developped an EM algorithm to infer the latent structure as well as the network in an elegant complete likelihood framework. The great advantage of this all-encompassing method was that the parameters of the latent clustering should guide the choice of regularization level. However, in practice the tuning parameter provided in the E-step (inference of the structure) were far too strong for the M-step (inference of the network). An other work by Marlin et al. (2009) provides a refined Bayesian algorithm to implement this approach. In the following, we adopt a fast and straightforward two-step approach illustrated in Figure 2.4: 1) definition of the latent structure $Z$ and corresponding penalty matrix $\mathbf{P}^{\mathbf{Z}}$, 2) structure adaptive inference of the network. Details about those two steps are given in the following sections.

### 2.3.1 Structure Inference

Coming up with a graphical topology, in other words a clustering of the genes informative with respect to the position of edges, can be realized in at least two ways. The first one is to find biologically grounded elements of structure based upon various computational biology or bibliographic tools. The second is to infer the latent structure using mixture models for graphs on an initial graph estimate $\widehat{\Gamma}_0$.

#### Examples Where to Find Biological Information about Latent Structure

Many sources can be used as prior biological structures, as long as they provide information on the pattern of regulations. We only provide here some hints at what could be useful. However these expert-based topologies highly depend on the biological model under study and the extent of expert knowledge available at the time of research.

A first source of information lies within metabolic pathways as available from the `KEGG` or `BioCarta` databases. Genes belonging to the same pathway are more likely to interact together and be connected in the regulatory network. This option was explored by Jeanmougin et al. (2011). When possible, information on which genes in the dataset code for transcription factors is highly relevant, particularly for time-course data.

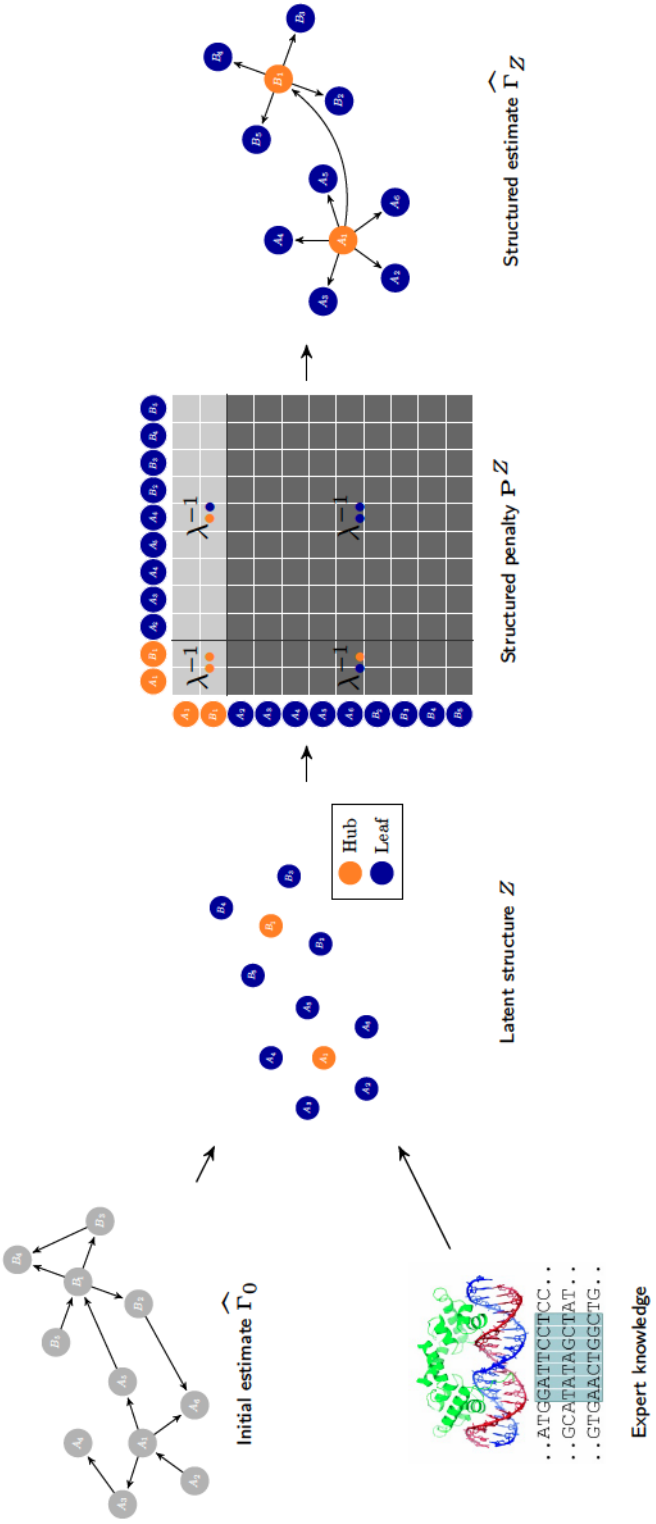



Figure 2.4 – *Overall inference strategy in two steps: 1) Definition of the latent structure by a) collecting prior information on the structure either by initial inference of the network via a usual $\ell_1$ regularized GGM or expert knowledge, b) defining the clustering of genes based upon step 1-a, c) designing the structured penalty matrix, 2) Structure adative inference of the GGM.*

When they exist, computational predictions of the number of potential binding sites for every known transcription factor in the data set is of even greater use to indicate where to look for potential edges. Such information is for instance available for *S. cerevisiae* in the `Yeastract` database.

**Statistical Inference of the Latent Structure**

An interesting graph modeling which captures the features of biological networks is the Stochastic Bloc Model (SBM) framework, providing mixture models for random graphs. This model has been rediscovered many times in the literature and a non exhaustive bibliography should include Frank and Harary (1982), Snijders and Nowicki (1997), Nowicki and Snijders (2001), Tallberg (2005), Daudin et al. (2008), Mariadassou and Robin (2007). The most important parameter, allowing to describe a large panel of network topologies, is the connectivity matrix $\boldsymbol{\pi} = (\pi_{q\ell})_{q,\ell \in \mathcal{Q}}$, describing $\mathbb{P}(i \leftrightarrow j | i \in q, j \in \ell)$, that is, how genes from each cluster connect to each others. Note that even though SBM models describe how to generate edges conditional on the clustering of genes, the use of SBM models follows the reverse path: the objective is to recover the clustering and connectivity coefficients which best fit the observed network.

Inference of such models, including directed SBM, has been implemented in various `R` packages, for instance `mixer` which is of straightforward use. Details about a large panel of methods to infer SBM can be found for instance in Daudin et al. (2008), Latouche et al. (2011).

SBM structures are integrated within step 1 by first, inferring an initial estimate $\widehat{A}_0$ and its corresponding graph $\widehat{\Gamma}_0$ based upon a usual unweighted $\ell_1$ penalty as in Problem 1.11 (step 1a in Figure 2.4); secondly, inferring the latent structure via an SBM algorithm on $\widehat{\Gamma}_0$ (step 1b); finally deriving the structured penalty matrix from SBM parameters (step 1c). Various penalty values can be defined as decreasing functions of the estimated connectivity matrix $\widehat{\boldsymbol{\pi}}$ . Suppose genes $g$ and $h$ are assigned to their most probable clusters $q$ and $\ell$, then an efficient penalty weight for edge $\Theta_{gh}$ is $\lambda_{q\ell} = 1 - \pi_{q\ell}$.

Of course, given the wide variety of topologies offered by SBMs, inferring such models contains a risk of overfitting if the number of observations is too small. Instead, a much simpler yet more robust model can prove itself efficient in the time-course setup, namely restricting the modeling of structure to the identification of hubs. To this purpose we suggest a very intuitive path. A first matrix $\widehat{A}_0$ is estimated using an adequate single Lasso penalty. Genes are then classified into two groups, hubs and leaves, according to the values of the $\ell_1$-norms of the corresponding rows in $\widehat{A}_0$. In order to account for the particularly strong heterogeneity between the two groups (differences in size and dispersion), our advice is to rely on a Gaussian mixture model to obtain the partition of genes between the two groups. This defines two submatrices $\widehat{A}_0^1$ and $\widehat{A}_0^2$ containing respectively the lines corresponding to the first and second groups. Hubs are then characterized as the class with the maximum mean absolute value of $\widehat{A}_0^k$.

### 2.3.2 Exact Neighborhood Selection for Network Inference

Once the internal structure has been recovered, inference of $A$ amounts to optimizing the penalized likelihood (2.7) where $\mathbf{Z}$ are fixed parameters. This can be achieved by solving some $p$ independent weighted Lasso problems in a neighborhood selection spirit (Meinshausen and Bühlmann 2006). Since there is no symmetry constraint on $A$, in this particular case and contrary to the i.i.d. setting exposed in Chapter 1, the neighborhood selection approach is exactly equivalent to maximizing the regularized likelihood of Equation (2.7). Details about the definition of an active-set algorithm adopting this approach are specified in Charbonnier et al. (2010).

**Remark 2.2** *With this approach, the sparsity constraint only applies to each column of $A$. This constraint implies that if we use $n+1$ time points, $\mathbf{S}$ is of rank $n$ and thus no more than $n$ connections can be activated by the Lasso at most in each column (assuming the penalty is low enough to accept the activation of all possible edges). Consequently, the sparsity constraint only applies to* incoming *edges and not to* outgoing *ones. In that sense, sparsity assumptions implied by $\ell_1$ penalization only assume that each node is regulated by a small set of nodes and do not contradict the existence of hubs regulating a huge set of nodes.*

## 2.4 Experiments and Discussion

In this section we apply our algorithm to both synthetic and real data. Comparison is made first within the family of the weighted-Lasso. We observe the performances of the Lasso when associated with a single Lasso penalty or an adaptive penalty. For the adaptive-Lasso, a single Lasso penalty is used as initial estimator. We then try two different hub penalties: one relying only on the known hub structure and another one inferring the hub structure from the initial Lasso estimator. We denote these estimators by *Lasso*, *Adaptive*, *KnwCl*, and *InfCl* respectively. Corresponding penalties can be summarized as follows:

$$\begin{aligned} P_{ij}^{\text{Lasso}} &\propto 1 \\ P_{ij}^{\text{Adaptive}} &\propto \left( \frac{1}{\hat{A}_{ij}^{\text{init}}} \vee 1 \right) \\ P_{ij}^{\text{KnwCl}} &\propto \left( \lambda_{\text{hub/leaf}} Z_{i,\text{hub}} + Z_{i,\text{leaf}} \right) \\ P_{ij}^{\text{InfCl}} &\propto \left( \lambda_{\text{hub/leaf}} \hat{Z}_{i,\text{hub}} + \hat{Z}_{i,\text{leaf}} \right), \end{aligned}$$

where $x \vee y = \max\{x, y\}$ and $\hat{Z}$ denotes the inferred classification. In the remainder of this section, we fix the ratio $\lambda_{\text{hub/leaf}} = 2$, thus penalizing twice as much nodes labeled as leaves as nodes labeled as hubs. Note also that we choose to maintain the modification of adaptive weights adopted in Zhou et al. (2009) and prevent the alleviation of penalty parameters. This trick ensures that the adaptive-Lasso will select a subnetwork from the network inferred by the initial Lasso estimate. No edge can be included if it was already excluded by the Lasso. In this way, the adaptive-Lasso guarantees a decrease in false positives.

Apart from our family of weighted-Lasso proposals, comparison will be made with state-of-the art network inference methods in a *VAR*1 setting: the *Shrinkage* method suggested by Opgen-Rhein and Strimmer (2007), the Recursive Elastic Net method (*Renet-VAR*) developed by Shimamura et al. (2009) and the method based on dynamic Bayesian networks proposed by Lèbre (2009), available in `R` within the `G1DBN` package.

Here, the interest of the inference lies in the recovery of the true edges, in other words of whether the entries of $A$ are correctly identified as nonzero. Our estimators are mainly used for discriminating nonzero entries from others. Quantities such as True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) summarize the performances of these classifiers. Precision TP/(TP + FP) is the ratio of the number of true nonzero elements to the total number of nonzero elements in the estimated matrix $\hat{A}$. Recall TP/(TP + FN) denotes the proportion of nonzero elements in $A$ which were correctly recovered as nonzero in the estimation, which corresponds to the usual statistical notion of power. Fallout FP/(FP+TN) gives on the contrary the proportion of zero elements in $A$ which were falsely declared as nonzero in the estimation. In statistical terms, the Recall (or Hit Rate) would be the empirical equivalent of the power of our classification method considered as a test, while the Fallout (or False Alarm Rate) would correspond to the first type $\alpha$ error. Note that, in the context of sparse network inference, the number of total positives is small compared to the number of total negatives. Thus, small variations of FP and TP will induce small variations in Fallout and large variations in Recall. Hence, comparison between Precision and Recall is generally more relevant than Fallout / Recall comparison in the present sparse context. This is why we will generally choose to omit Fallout rates when we need to alleviate the presentation of results.

These rates are easily obtained for the Lasso based methods since they automatically produce null coefficients. By increasing the penalty parameter we obtain sparser and sparser graphs. We start from a large enough penalty to constrain all coefficients of $\hat{A}$ to 0 and decrease the penalty until we include as many variables as allowed by the ratio $n/p$. We then select the best penalty from this list as the one maximizing either the BIC or the AIC criterion.

Like the Lasso, *Renet-VAR* directly implements variable selection and penalty choice is included in the algorithm. Concerning *G1DBN*, we follow the author's advice to tune the parameters of the test procedure as described in the additional material of Lèbre (2009). When applying the *Shrinkage* method developped by Opgen-Rhein and Strimmer (2007), a supplementary step is required to transform continuous results into a binary solution. We follow Opgen-Rhein and Strimmer's advice and rely on local false discovery rates. This provides each edge with an existence probability conditional on the corresponding entry in $\hat{A}$. We declare as inferred edge any edge with posterior probability exceeding the threshold of 80% as the authors do.

### 2.4.1 Simulated Data

**Simulation Settings.** To assess the performances of our approach, we apply the $VAR1$ model to a very favorable setup, where existing models already perform quite well. We then decrease the ratio $n/p$ in order to observe the response of each method to this increasing lack of information. On top of that, we consider graphs of different sizes: small graphs of 20 nodes, larger graphs of 100 nodes and a setup with 800 nodes. For smaller graphs, we consider three different amounts of observations: 10, 20 and 40. For medium sized graphs, we also consider the cases $n = p/2$ and $n = p$ but omit the case $n = 2p$ as unrealistic. The setup $p = 800, n = 20$ is meant to mimic Spellman et al. (1998)'s dataset.

Simulation of the $VAR1$ process is based upon the simulation strategy used by Opgen-Rhein and Strimmer (2007) in order to ease the comparisons, but introduces a structure based on hubs in order to better reflect the structure we could expect from a real data set. A graph is first simulated, with fixed numbers of nodes and edges. Like Opgen-Rhein and Strimmer (2007) we simulate sparse graphs, with $K = 2p$ edges. Nodes are split into two groups according to a multinomial distribution with probabilities (0.1,0.9), leading to 10% of hubs in average. Edges are then positioned in the graph according to a multinomial distribution, with 85% of edges from hubs to leafs, and the remaining set within hubs. Exception is made for the very large graph, for which we base the number of edges and their distribution on Spellman et al. (1998)'s data. The matrix $A$ is synthesized on the basis of this graph: we attribute a random partial correlation value uniformly distributed on $[-1, -0.2] \cup [0.2, 1]$ to all nonzero coefficients (corresponding to edges in the graph).

From this matrix, a $VAR1$ observation is generated, using a centered Gaussian starting value and a centered Gaussian noise, both with variance $\sigma^2 = 0.1$. For computing time reasons, this is repeated 500 times for the small graphs, 200 times for medium sized graphs and 100 times for the large graph. Results are averaged over all samples.

To gain a better insight into the difficulty of these synthesized datasets for a Lasso estimator, we checked whether the *irrepresentability condition* (Zhao and Yu 2006, Meinshausen and Yu 2009) recalled in Chapter 1 was validated in all these very simple simulations. First, note that the graphical context requires the irrepresentability condition to be validated for each of the $p$ genes at the same time, which makes it much more difficult to hold than in the simple regression context where it is an already strong hypothesis. In our context, since we solve $p$ independent Lasso problems, we can check the validity of the hypothesis in each of these individual problems. For each gene, the irrepresentability condition is tested using the true sign pattern extracted from the corresponding column of the true adjacency matrix. Thus the sets of relevant and irrelevant covariates are allowed to vary from one problem to another. Generating 100 samples of each simulation setting, we observed that even in a favorable setup with twice as many observations as variables ($p = 20$ genes) the irrepresentability condition fails for 30% of genes in average. With $p = 20$ genes and only $n = 10$ observations this assumption fails on average for 51% of the genes.

In other words, for around half of the genes we cannot expect the Lasso to recover the exact sign pattern. See Table 2.1 for details. Admittedly, the irrepresentability condition is a really strong assumption, necessary and sufficient for exact sign recovery, that is to say not only the exact neighborhoods (no false positives, no false negatives) but also the exact signs of the correlations. Yet since the simulated values are quite well separated between true zeros and true nonzeros we would have expected that this hypothesis would have been much more validated. Information about the validity of the *restricted eigen-value assumptions* (Bickel et al. 2009) would be greatly appreciated to compensate for such pessimistic results, but these are computationally intractable. Adaptation of Juditsky and Nemirovsky (2008)'s results to the present context could be of great benefit.

| $n/p \diagdown p$ | 20 | 100 |
|---|---|---|
| 2 | 0.30 (0.23) | - |
| 1 | 0.41 (0.23) | 0.37 (0.15) |
| 1/2 | 0.51 (0.18) | 0.42 (0.12) |

Table 2.1 – *Average proportion of genes for which the irrepresentability condition does not hold and standard error in each simulation setting (hence the empty cell for* $p = 100$, $n = 200$*).*

**Discussion of Simulation Results.** Results are presented in Figure 2.5 under the form of Barcharts. Figure 2.6 illustrates the case where $p = 100$ by giving boxplots for the distributions of Precision, Recall and Fallout.

Compared methods differ with the type of setting. First of all, since the *Shrinkage* method (particularly the local false discovery rate step) relies on the hypothesis that $p$ is large, we do not consider it fair to apply it to the small network setting. Reversely, for computing time reasons we decided to restrict the application of *G1DBN* to the graphs of size $p = 20$.

Penalties for the Lasso based methods were chosen on the basis of either the BIC or AIC criteria. Although theory states that the BIC ought to outperform the AIC in terms of model selection (Zou et al. 2007), we observed that in practice the BIC criterion might be too conservative when $n$ is small compared to $p$. In that situation, it might be interesting to favor the less stringent AIC criterion which will induce a higher recall rate for not such a large loss in precision. Note that the penalty choice based on the AIC or the BIC can lead to choose the null model as best model. In that case, Precision cannot be defined. We thus show the results for precision over all simulations where at least one variable was included.

The first point worth noting in Figure 2.5 is that in all settings the Lasso is outperformed by weighted-Lasso methods and others. This quick check confirms the interest of compensating for the bias induced by $\ell_1$ regularization on large coefficients. It is also possible that what we observe about the validity of the irrepresentability condition jeopardizes the performances of the single-penalty Lasso. In line with Table 2.1, the Lasso performs particularly badly when the ratio $n/p$ is not favorable, with recall and precision rates under 20% when $p = 20, n = 10$. It even performs so poorly that it deprecates the inference based on adaptive weights. Prior information on where the true zeros might compensate for this apparent lack of "neighborhood stability", using Meinshausen and Bühlmann's

vocabulary, and explain why the *KnwCl* penalty is far more accurate (precision of 84% in average for a recall of nearly 50% in average for the same simulation setting $p = 20, n = 10$).

As expected, in all settings (except when $n$ is really too small compared to $p$) the *Adaptive* penalty improves the precision but at the price of a smaller recall rate. On the contrary, the inferred classification *InfCl* allows to improve the precision without undermining the recall rate. However, both methods are highly dependent on the initial Lasso estimate. Therefore, the gain in precision resulting from such methods decreases with the $n/p$ ratio.

Benefitting from a certain amount of supplementary information, the *KnwCl* penalty leads to a clear increase in both precision and recall. Particularly when little information is available in terms of number of observations, taking prior information about which genes are potential regulators and which are not into account improves the results dramatically. This is true when compared to all Lasso based methods but generalizes to *Shrinkage*, *Renet-VAR* and *G1DBN*. Admittedly, *Renet-VAR* leads to higher precision values with medium sized graphs, but it is compensated by smaller recall rates.

Table 2.5 shows naturally that we cannot expect too much from very extreme settings ($p = 800, n = 20$, that is, the Spellman et al.'s settings). Average Recall rate is less than 20% for all methods except the *KnwCl* penalty. In this case, knowledge of potential hubs allows the recall rate to almost double in average while increasing the precision. Note however that even with this supplementary information precision rates never exceed 50%.

To finish with, we would like to lay the emphasis on computing times. For this we let the number of nodes range from 5 to 185 and fixed the number of observations at half the maximum number of nodes, i.e. $n = 92$. This leads to a ratio $n/p$ ranging from 0.05 to 2. Computing times for the weighted-Lasso with inference of the classification *InfCl* and selection of the best penalty, *Renet-VAR* and *G1DBN* are presented in the log-log scale in Figure 2.7. We can see that running times for *Renet-VAR* and *G1DBN* can become a handicap as soon as $p$ gets large while computing times for *InfCl* rarely exceed 2 minutes.

### 2.4.2 Yeast Data

We confronted our model to time measurements of *Saccharomyces cerevisiae* gene expression data collected by Spellman et al. (1998). We focus on the subset of genes they identified as periodic, i.e. genes whose transcription levels over time show evidence that they are cell-cycle regulated.

**Remarks on the Data Set.** This dataset is one of the first microarray experiments. It is thus doomed to be rather noisy, contrary to the simulated data sets. Besides, we had to face the problem of missing values, which appeared on some of the most important genes. We imputed them as the mean of the two closer known observations in time for the gene considered, before and after the time point of interest.

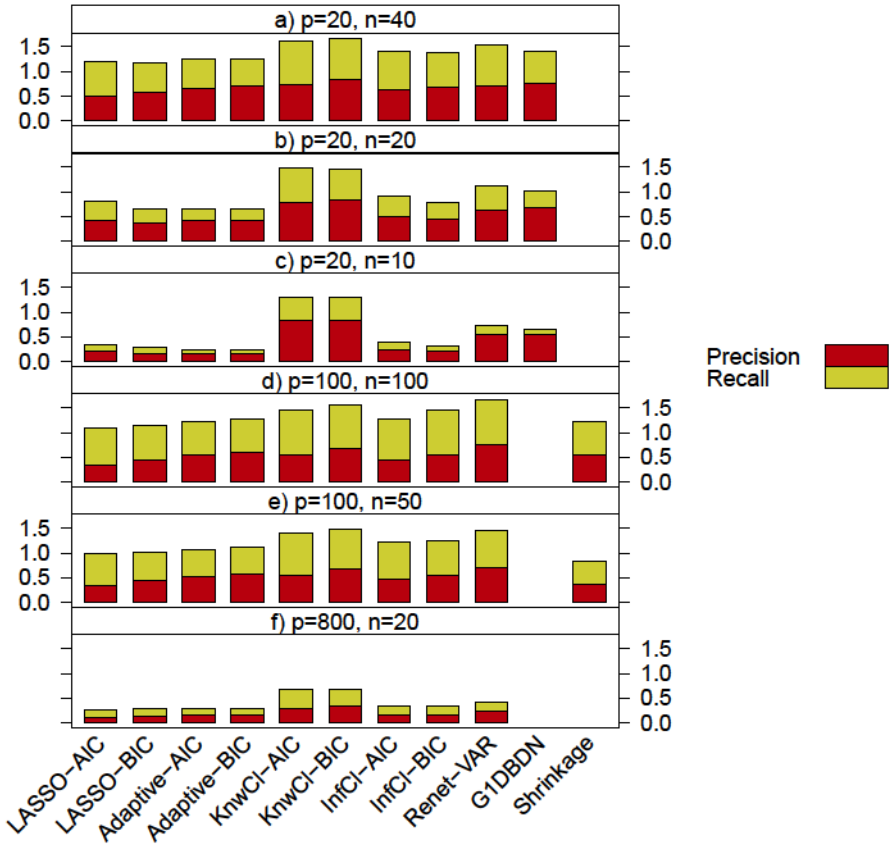



Figure 2.5 – *Bar charts of Precision and Recall rates for each method and simulation setting, averaged over all simulation samples.*

On top of its noisiness, Spellman et al.'s data set is particularly hard to tackle from a statistical view point. Information is provided on 786 genes for only 18 time points. This implies that using our algorithm we cannot activate more than $17 * 786 = 13362$ edges out of $789 * 786 = 617796$ possible ones, that is to say 2.2%.

However, we can rely on experimental conclusions on yeast gene regulation networks to collect target information about the true edges of the graph. We compare our results to the adjacency matrix provided by the Yeastract database (`www.yeastract.com`). We retain information on documented direct relationships, that is to say direct regulations confirmed by published experimental results.

Note however that this theoretical benchmark is biased in two ways. First, some true edges might be missing because all regulations might not have been confirmed by experiments yet. Second, this graph gathers all reported regulations, whatever the conditions of the experiment. Some might not actually happen during the precise experiment we consider. We can suppose the effect of the first bias to be low in a model organism such as *Saccharomyces cerevisiae*. The effect of the second bias is much more likely however, since measurements are all made while cells are at the beginning of their growth, growing until ready for DNA synthesis. We

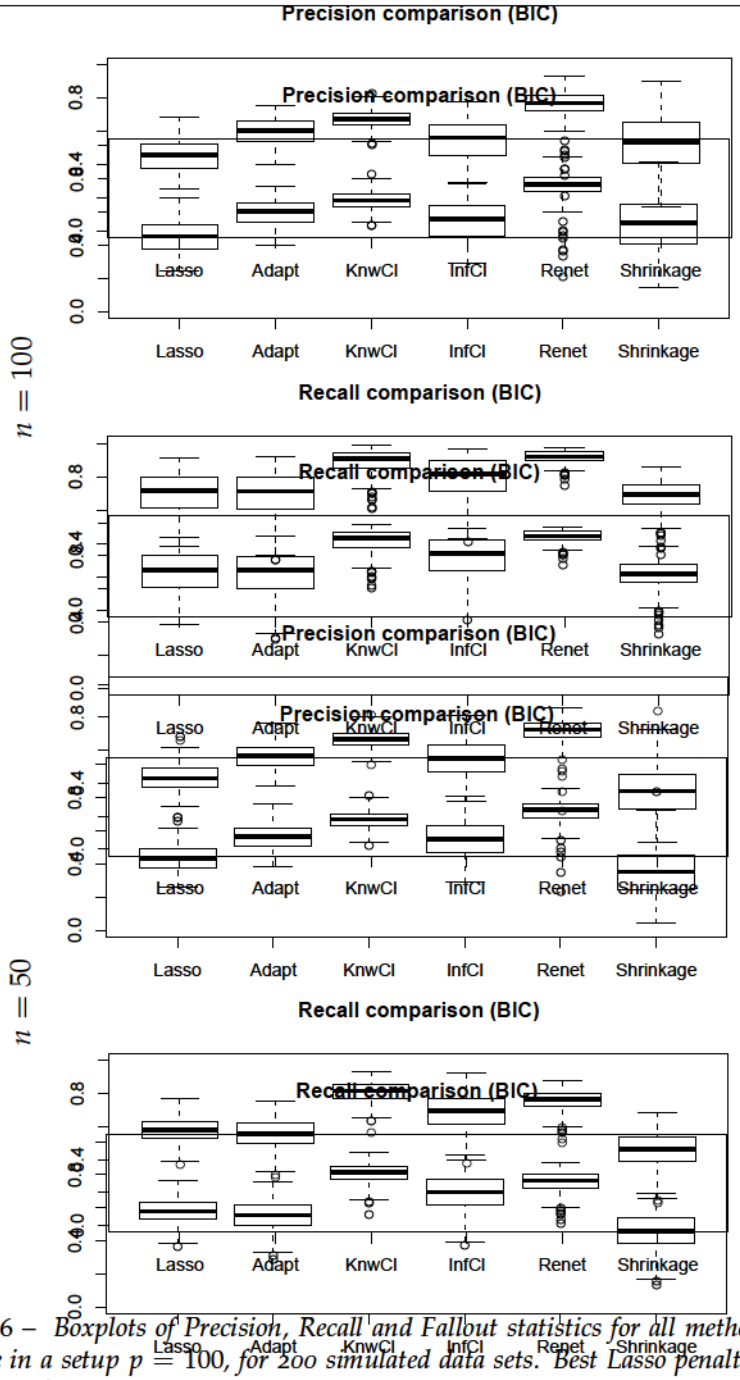



Figure 2.6 – *Boxplots of Precision, Recall and Fallout statistics for all methods except Shrinkage in a setup $p = 100$, for 200 simulated data sets. Best Lasso penalties chosen on the basis of the BIC criterion.*

cannot expect the whole range of possible regulations to happen in such a small portion of the cell cycle.

This dataset illustrates quite well the biological properties our model is based upon. First, documented information reveals the existence of 1385 true edges (among more than 600000 possible ones in theory). The theoretical graph is thus extremely sparse. Secondly, the hub structure is quite clear: edges leave from only 26 out of 786 genes. Hence knowledge of the hubs provides crucial information on the position of edges. This phenomenon also clearly appears on Figure 2.8. Incoming degrees never exceed 20 but only 1 is null. On the contrary, outgoing degrees are null for the vast majority of genes. Significant degrees appear as outliers in this distribution, reaching up to 150 for some of them.

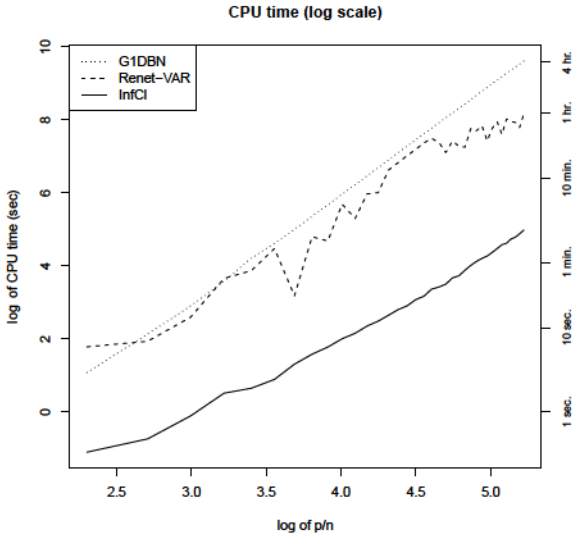



Figure 2.7 – *Computing times on the log-log scale for Renet-VAR, G1DBN and InfCl (including inference of classes). Intel Dual Core 3.40 GHz processor.*

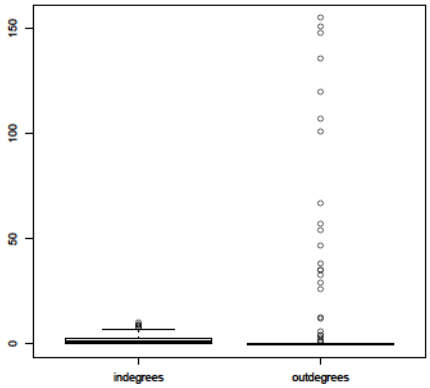



Figure 2.8 – *Boxplots of incoming and outgoing degrees in Yeast theoretical adjacency matrix*

**Discussion of the Results.** The setting is much harder than in the first simulated data sets, with a ratio $n/p = 2.3\%$ as well as harder than the last simulated dataset with less separated correlations between existing and non existing edges. Results presented in Table 2.2 show quite well the difficulty all methods encounter in front of this data set. Results for the *Shrinkage* approach are not shown because the local false discovery rate step included in this method was heavily flawed by the lack of separability between edges and non edges. Except for the *KnwCl* penalty, all Lasso based estimators are reduced to the null model. Both the BIC and AIC criteria do not find the increase in likelihood large enough to compensate for the complexity of any model with at least one edge. Performances of the *KnwCl* penalty and *Renet-VAR* remain lower than what we could expect from simulated results.

Many reasons for such bad perfomances could be thought of. We already mentionned the noisiness of the data, which quite hardly differentiated the edges from non edges. Second, homogeneity of the VAR1 model might be too strong an assumption. Last but not least, when looking more

| Models | Lasso | Adaptive | KnownCl | InferCl | Renet |
|---|---|---|---|---|---|
| Precision | - | - | 0.082 | - | 0.004 |
| Recall | 0 | 0 | 0.068 | 0 | 0.003 |
| Fallout | 0 | 0 | 0.002 | 0 | 0.002 |

Table 2.2 – *Precision, Recall and Fallout performances for all Lasso based methods and Renet-VAR on Spellman et al.'s data set. Best Lasso penalties chosen on the basis of the BIC criterion.*

closely at how data were collected we noticed that measurements were made every 7 minutes, which might be long enough for dependencies to vanish. Also, since we measure values related to the cell cycle, measurements were necessarily made on different cells each time, thus measuring the expression levels on different individuals at each time point. In brief, this apparently longitudinal data set might share more common points with i.i.d. models than with VAR1 processes.

### 2.4.3 E. coli S.O.S. DNA Repair Network

In this section we quit the high dimensional setup and compare the performances of all methods in a much easier framework. We focus on a sub-network from *E. Coli* S.O.S. DNA repair network analyzed by Ronen et al. (2002) [1]. Data provide information on the main 8 genes of the S.O.S. network (*uvrD,lexA,umuD,recA,uvrA,uvrY,ruvA* and *polB*) across 50 time points. Measurements rely on precise expression kinetics which allow Ronen et al. (2002) to monitor mRNA expression levels every 6 minutes after exposition of the DNA to UV light at time 0. We will not dwell on the measurement technology here (see Ronen et al. 2002, for details). Note however that the authors do not measure the actual mRNA quantity present in the cell at time $t$ but the instant promoter activity of each gene. Equivalence between the two measurements is guaranteed if the instant quantity of mRNA in the cell roughly equals its production rate, that is to say if there is no accumulation of mRNA in the cell. Under this assumption, Ronen et al. (2002) 's data can be used as any microarray dataset.

*E. coli* S.O.S. DNA repair network provides a precise benchmark: specific regulatory interactions in response to DNA damage have been characterized. In other words, we can rely on a theoretical regulatory network which represents the main direct transcriptory regulations actually taking place during the experiment. According to the regularly updated EcoCyc database, *lexA* is the only regulator in this subnetwork, regulating all genes including itself. Concretly, the protein *LexA* is at the core of the regulation network, usually binding sites in the promoter regions of S.O.S. genes to repress their expression. As soon as *RecA* senses DNA damage (by binding to single-stranded DNA), it becomes activated and induces *LexA* autocleavage. The decrease in *LexA* concentration alleviates the repression of S.O.S. genes. When damage is repaired, the level of activated *RecA* drops, *LexA* accumulates and represses again all S.O.S. genes.

Detailed results are presented in Figure 2.9. We can see that performances differ a lot from one experiment to another. Particularly, experi-

[1] data downloadable on Uri Alon's homepage, `http://www.weizmann.ac.il/mcb/UriAlon/`

ments 1 and 4 lead to significantly poor results although nothing should *a priori* distinguish them from 2 or 3 (1 and 2, respectively 3 and 4, share the same U.V. exposure).

As on simulated data, the Lasso leads to poor results. *G1DBN* shows similarly poor performances here. Quite surprisingly, *Renet-VAR* does not perform as well as we could have expected from simulations. It reaches 50% of recall at the expense of very low precision rates. *Adaptive* penalty improves more the quality of the estimation than in the simulation studies. Now they increase the precision of the Lasso without really undermining the recall rate. Inference of the classification outperforms these, with higher recall and precision rates. This is quite interesting since except in experiments 1 and 4 where the Lasso provide almost no information, inference of the classes seems quite good although the initial Lasso still shows mediocre results. To finish with, the *KnwCl* penalty benefits quite well here from its extra information since it outperforms all other methods and manages to reach honest results even in datasets 1 and 4 which disturbed all other methods.

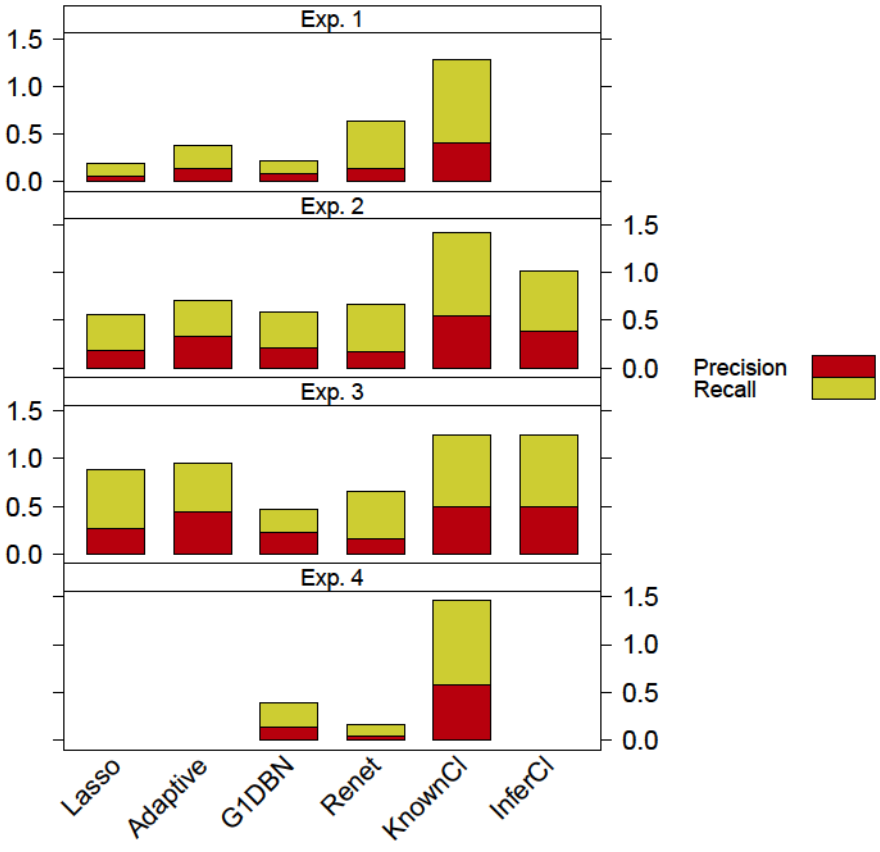



Figure 2.9 – *Bar charts of Precision and Recall rates for each method and experiment.*

Inferred graphs on experiment 2 are shown in Figure 2.10. The regulatory activity of *lexA* is more or less recovered by all methods. What is interesting is that a common structure recurently shows up among false positives: regulations due to *uvrA*. This regulation pattern is particularly

what dominates experiment 4 and leads to so poor results. Strangely, we could not find any mention of this regulatory activity in the literature. Either there is a need for further biological research on this gene or there is an undirect regulation blurring the results. Another unknown regulation dominates all inferred graphs: regulation of *uvrY* by *polB*. It is all the more interesting as it survives the bad *a priori* that the *KnwCl* penalty holds against it. Further biological investigation could want to look at this couple of genes more closely.

In this respect, we could note that the regulatory effect of activated *RecA* on *LexA* does not appear on these graphs, which we could see as a good point since this is a post-transcriptional regulation. We would also like to lay the emphasis on the fact that we here check selection consistency of all the methods but not their sign consistency. We only check whether we identify the right edges and not the activation/inhibition processes associated to them. Looking more closely at the estimated matrices, we can see that the (shrunk) correlations estimated between *lexA* and the remaining genes are all positive and not negative as the literature would tell. This would not be a flaw in all methods but a direct result of the limitations of transcriptomic data. Indeed, we only observe mRNA production rates. As a consequence, we cannot spot the decrease in concentration of protein *LexA* and only observe that the expression of all genes suddenly increases, *lexA* included.

## 2.5 Conclusion

This chapter presents a weighted-Lasso algorithm designed to tackle time varying gene expression data taking into account an underlying structure. In this particular framework, the proposed approach outperforms similar methods. Even when regulators and regulatees cannot *a priori* been distinguished through analysis of the literature, inference of the classification greatly improves the performances of the Lasso. It therefore seems good to advice that, whenever available, knowledge about potential transcription factors should be taken into account and that basic knowledge on the topology of biological networks should not be omitted in the modeling process.

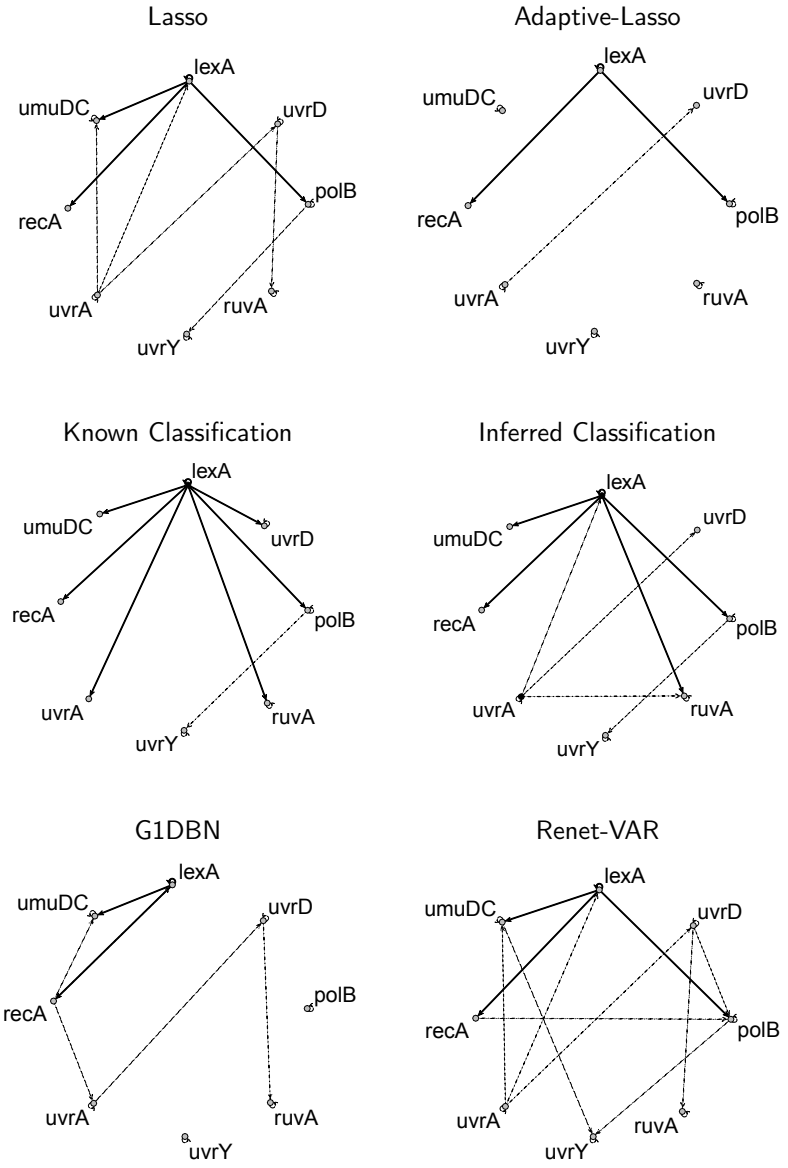



Figure 2.10 – *Graphs inferred by the different methods on experiment 2 data. Lasso penalties are chosen so as to maximize the BIC criterion. True positives are drawn in black while false positives are shown in dashed gray.*

# Consistency analysis of the cooperative-Lasso

3

The cooperative-Lasso was introduced by Chiquet et al. (2011) in the context of multiple Gaussian graphical models. More generally, the cooperative-Lasso tackles the issues of estimation and selection of parameters endowed with a known group structure, when the groups are assumed to be *sign-coherent*.
The present chapter sheds light on the derivation of optimality conditions and consistency properties of this new regularization. We prove asymptotic consistency in terms of model selection and non-asymptotic oracle inequalites in terms of estimation and prediction.
Finally, we provide an illustration of the benefits of the cooperative-Lasso in the context of multiple Gaussian graphical model inference on a longitudinal treatment/placebo experiment in Multiple Sclerosis and on a case/control study in Breast Cancer.
The asymptotic model selection property has been published as part of Chiquet et al. (2012), but the remaining of the chapter presents new results.

Contents

## Introduction

The idea of the *cooperative-Lasso* originates in the inference of joint Gaussian graphical models from distinct but related transcriptomic datasets, as introduced in Chiquet et al. (2011). Indeed, many transcriptomic experiments are led simultaneously in many *close* conditions, as part of a more general experimental scheme. Stress experiments, case/control studies, placebo/treatment studies form a non-exhaustive list of such multiple condition experiments.

Multiple Gaussian graphical models are one among a growing list of applications where several datasets or covariates within the same dataset provide independant but redundant information about a single statistical problem: multichannel signals, video denoising or inpainting, multiple response problems (Turlach et al. 2005), gene signatures based upon clusters of co-expressed genes (Eisen et al. 1998, Park et al. 2006, Ma et al. 2007), etc. In this configuration, there is a clear gain in combining information from all datasets in a refined way, compared to the treatment of each dataset independently from the others. In the same spirit, there is an increasing interest in combining information provided by highly correlated or somewhat redudant variables, instead of letting them compete against each other in the quest for *the* unique best model.

These examples illustrate many settings where there is much more interest in considering subsets of covariates jointly rather than letting them compete in a falsely independent framework. When the same statistical model is considered in close but distinct experiments, one might be interested in keeping the inference of distinct parameters, but joining information across experiments in order to add robustness to the selection of relevant features. When datasets present lots of highly correlated covariates, one may want to avoid working on a reduced subset of uncorrelated but arbitrarily chosen covariates, and work instead on the full set of covariates, taking advantage from the redundancies in place.

We adopt the linear regression model as main framework for theoretical developments. Assume we observe a continuous response variable $Y$ that we want to predict from a vector of $p$ predictor variables $X = (X_1, \dots, X_p)$ partitioned into $K$ groups $\{\mathcal{G}_k\}_{k=1}^K$ of respective sizes $\{p_k\}_{k=1}^K$. Covariates belonging to the same group present some redundance of information, be it dictated *a priori* by the experimental design (same covariates in multiple conditions, multichannel signals, several probes related to the same gene, ...) or be it decided *a posteriori* because covariates are too correlated with each other to be treated independently. We work on the following model:

$$Y = X\boldsymbol{\beta}^\star + \varepsilon = \begin{bmatrix} X_{\mathcal{G}_1} & X_{\mathcal{G}_2} & \dots & X_{\mathcal{G}_K} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^\star_{\mathcal{G}_1} \\ \boldsymbol{\beta}^\star_{\mathcal{G}_2} \\ \vdots \\ \boldsymbol{\beta}^\star_{\mathcal{G}_K} \end{bmatrix} + \varepsilon \ , \tag{3.1}$$

The error term $\varepsilon$ is assumed zero-mean Gaussian with variance $\sigma^2$. The estimation of $\boldsymbol{\beta}^\star$ is based on the vector $\mathbf{y} = (y_1, \dots, y_n)^\intercal$ of responses and an $n \times p$ design matrix $\mathbf{X}$ whose $j$th column contains $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})^\intercal$,

the $n$ observations for variable $X_j$. For clarity, we assume that both $\mathbf{y}$ and $\{\mathbf{x}_j\}_{j=1,\dots,p}$ are centered so as to eliminate the intercept from fitting criteria. Multitask datasets easily fall into this framework, as will be illustrated on joint Gaussian graphical models in Section 3.4.

Many regularization terms have been suggested in the recent years in order to meet at the same time the requirements of high-dimensional inference and joint sparsity patterns. Most of them are based on mixed $\ell_{1,\alpha}$ norms, $\alpha > 1$, the $\ell_1$ norm acting as a selection tool at the group-level embedding a group-specific $\ell_\alpha$ norm which breaks the independence of covariates within groups. Among those, $\ell_{1,2}$ and $\ell_{1,\infty}$ are certainly the most popular.

Independently proposed by Grandvalet and Canu (1999) and Bakin (1999) and later developed by Yuan and Lin (2006), the mixed $\ell_{1,2}$ regularization is often referred to as the group-Lasso penalty or block $\ell_{1,2}$ regularization, henceforth equivalently referred to as $\text{pen}_{\text{group}}(\cdot)$ or $\|\cdot\|_{1,2}$:

$$\text{pen}_{\text{group}}(\boldsymbol{\beta}) = \sum_{k=1}^{K} w_k \, \|\boldsymbol{\beta}_{\mathcal{G}_k}\|_2.$$

Weights $w_k > 0$ adapt the level of penalty within a given group. Typically, one sets $w_k = \sqrt{p_k}$, where $p_k$ is the cardinality of $\mathcal{G}_k$ in order to adjust shrinkage according to group sizes.

Following the propositions by Turlach et al. (2005), Tropp et al. (2006) there has also been an appeal for block $\ell_{1,\infty}$ regularizations.

$$\text{pen}_{\text{block}}(\boldsymbol{\beta}) = \sum_{k=1}^{K} \|\boldsymbol{\beta}_{\mathcal{G}_k}\|_\infty = \sum_{k=1}^{K} \max_{j \in \mathcal{G}_k} |\boldsymbol{\beta}_j|.$$

However, these regularizations completely condition the selection of a covariate to the selection of all other covariates within its group. Negahban and Wainwright (2011) and Huang and Zhang (2010) point out the perils of block-regularizations which suffer from deteriorated performances compared to a simple Lasso when sparsity patterns within groups get blurred. In order to overcome this restriction, simulatenous suggestions were made to combine mixed $\ell_{1,\alpha}$ norms together with an $\ell_1$ penalty, in the vein of the hierarchical penalties of Zhao et al. (2009). This is the case of the sparse group-Lasso by Friedman et al. (2010) or of dirty multi-task learning by Jalali et al. (2010; 2011):

$$\begin{aligned}
\text{pen}_{\text{spg}}(\boldsymbol{\beta}) &= \alpha \, \text{pen}_{\text{group}}(\boldsymbol{\beta}) + (1-\alpha) \, \text{pen}_{\ell_1}(\boldsymbol{\beta}); \\
\text{pen}_{\text{dirty}}(\boldsymbol{\beta}) &= \alpha \, \text{pen}_{\text{block}}(\boldsymbol{\beta}) + (1-\alpha) \, \text{pen}_{\ell_1}(\boldsymbol{\beta}).
\end{aligned}$$

The sparse group-Lasso or dirty modeling provide additional flexibility to the selection of covariates within groups but demand an additional tuning parameter $\alpha$. Chiquet et al. (2011) introduce a novel penalty that takes a different stance, with the benefit of requiring a single tuning parameter. The cooperative-Lasso, in short coop-Lasso, performs a sign-adaptive selection of grouped variables, dissociating the activation of positive and negative coefficients. Let $u^+$ and $u^-$ denote respectively the positive and negative parts of any vector $u \in \mathbb{R}^p$. The cooperative penalty

is defined as the sum of the group norms of the positive and negative parts of $\boldsymbol{\beta}$:

$$\text{pen}_{\text{coop}}(\boldsymbol{\beta}) = \text{pen}_{\text{group}}(\boldsymbol{\beta}^+) + \text{pen}_{\text{group}}(\boldsymbol{\beta}^-).$$

This strategy presents the advantage of being in full adequacy with the experimental design of various statistical problems, where groups are most likely to be sign-coherent. Chiquet et al. (2012) describes three applications where sign-coherence is a sensible assumption. The first one considers ordered categorical data, which are common in regression and classification. The coop-Lasso can be used to induce a monotonic response to the ordered levels of a covariate, without translating each level of the categorical variable into a prescribed quantitative value. The second application describes the situation where redundancy in probe measurements related to a same gene causes sign-coherence to be expected. Similar behaviors should be observed when features have been grouped by a clustering algorithm such as average linkage hierarchical clustering, which are nowadays routinely used for grouping genes in microarray data analysis (Eisen et al. 1998, Park et al. 2006, Ma et al. 2007). In Section 3.4 we focus on the inference of joint Gaussian graphical models, where underlying biological mechanisms make it sensible to assume that up- or down-regulations can disappear in some conditions, but are a lot less likely to reverse from one to the other.

Theoretical properties of $\ell_{1,2}$ and $\ell_{1,\infty}$ regularizations have been studied quite extensively in the past few years, following this enthusiasm for joint sparsity modeling. Bach (2008b) provides a first asymptotic analysis of the group-Lasso in terms of support recovery, under some irrepresentable condition or mutual incoherence assumption on the design, in the classical framework where the number of variables $p$ is fixed while the sample size $n$ grows to infinity. The first results to fit high-dimensional settings are to be seen in Meier et al. (2008), which derives bounds on the group-Lasso prediction error in generalized linear models, and Nardi and Rinaldo (2008), which derives bounds on the group-Lasso prediction and estimation error in linear regression. Lounici et al. (2009) provides sharpest sparsity oracle inequalities for prediction and estimation errors under restricted eigen-value assumptions in multi-task settings, shedding light on the advantage of block-regularization over the Lasso when sparsity patterns coincide with the group structure. In this chapter, we recall the classical asymptotic results published in Chiquet et al. (2012) validating model selection properties of the coop-Lasso, under less stringent assumptions than the group-Lasso. To answer the high-dimensional challenge, we add sparsity oracle inequalities for prediction and estimation errors in the spirit of Lounici et al. (2009), valid for any sample size and number of variables.

More recently, Lounici et al. (2011) have extended sparse oracle inequalities to linear regressions with grouped variables in general, while S. Negahban and Yu (2012) provide oracle inequalities under looser assumptions of sparsity. On top of that, high-dimensional model selection guarantees are provided by Obozinski et al. (2011), along with sample complexity functions, quantifying the reduction in sample size $n$ required for selection consistency, as a function of the number of variables $p$. Ne-

gahban and Wainwright (2011) provide similar results for $\ell_{1,\infty}$ regularization. How these state-of-the art results can be adapted to improve our theoretical analysis of the coop-Lasso will be addressed in discussion.

The first part of this chapter describes the coop-norm, as a sign-adaptive mixed-norm, in order to derive equivalence relationships and dual bounds which are essential to further theoretical developments. In the second part we derive optimality conditions for the coop-Lasso as well as one of its dual forms. Asymptotic model selection properties are recalled in a third section, along with sparsity oracle inequalities on prediction and estimation errors in the spirit of Lounici et al. (2009). An application of the coop-Lasso to the inference of joint Gaussian graphical models is detailed in the last part.

## 3.1 Cooperative Norms and Related Analysis Tools

We take a short detour in the analysis of our problem to study in more depth the properties of the cooperative norm. Not only some of those are required to prove oracle inequalities about the cooperative Lasso, but this also provides a better insight into this sign-adaptive group penalty. We start by recalling some properties of the group-Lasso penalty, considered as a mixed-norm, and derive similar properties for the cooperative norm and its variants. Most of these developments are linked to the central notion of dual norm $\|\cdot\|_\star$, defined for the norm $\|\cdot\|$ by:

$$\|x\|_\star = \sup_{\|y\|\leq 1} \langle x, y\rangle.$$

### 3.1.1 The Group-Lasso Penalty as a Mixed-Norm

The group-Lasso penalty is a special case of general mixed $\ell_{p,q}$ norms, namely with $p = 1$ and $q = 2$. Those norms were introduced in functional analysis and are now very popular in the inference of joint sparse problems in statistics, machine learning and signal processing (Zhao et al. 2009, Kowalski 2009, Szafranski et al. 2010, Obozinski et al. 2011). Besides the $\ell_{1,2}$ norm popularized by the group-Lasso, the $\ell_{1,\infty}$ norm has recently been the focus of increasing attention (Jalali et al. 2010; 2011, Negahban and Wainwright 2011). In the finite-dimensional case, mixed norms correspond to the composition of an $\ell_p$ and an $\ell_q$ norm on vectors equipped with double indices. Double indices naturally arise in many fields: for instance, indivuals and time-points in panel datasets, multichannel signals in signal processing,.... In the following, we define a hierarchy between those two labels, and refer to the first label as a group index.

Consider $p, q \in [1, \infty]$. The mixed $\ell_{p,q}$ norm associated to the group structure $\{\mathcal{G}_k\}_{k=1}^K$ is defined by the $\ell_p$ norm of $\ell_q$ norms on groups $\mathcal{G}_k$. For every $x \in \mathbb{R}^p$ with associated group structure $\{\mathcal{G}_k\}_{k=1}^K$, the $\ell_{p,q}$ norm reads:

$$\|x\|_{p,q} = \left(\sum_{k=1}^K \|x_{\mathcal{G}_k}\|_q^p\right)^{1/p} = \left(\sum_{k=1}^K \Big(\sum_{j\in\mathcal{G}_k} |x_j|^q\Big)^{p/q}\right)^{1/p}. \tag{3.2}$$

In the case where $p$ or $q$ are set to $\infty$, the corresponding norm is replaced by the supremum.

Observe from Equation (3.2) that for every $p > 0$, the mixed $\ell_{p,p}$ boils down to the usual $\ell_p$ norm. Equivalence relationships between mixed $\ell_{p,q}$ norms can be straightforwardly derived from usual equivalence relationships between $\ell_p$ norms. As a result of the composition of the $\ell_p$ norm onto the $\ell_q$ norm, the number of groups replaces the total number of components in the compatibility constant. For every $x$ and $y \in \mathbb{R}^p$, the mixed norms associated to the group structure $\{\mathcal{G}_k\}_{k=1}^K$ satisfy:

$$\begin{array}{rcccl} \|x\|_2 & \leq & \|x\|_{1,2} & \leq & \sqrt{K}\ \|x\|_2, \\ \|x\|_{\infty,2} & \leq & \|x\|_{1,2} & \leq & K\ \|x\|_{\infty,2}, \\ \|x\|_{\infty,2} & \leq & \|x\|_2 & \leq & \sqrt{K}\ \|x\|_{\infty,2}. \end{array}$$

Similarly, Hölder's inequality generalizes to mixed norms. For every $x$ and $y \in \mathbb{R}^p$, the mixed norms associated to the group structure $\{\mathcal{G}_k\}_{k=1}^K$ satisfy for every two pairs of conjugates $(p, q)$ and $(p', q')$ such that $1/p + 1/q = 1$ and $1/p' + 1/q' = 1$ the following Hölder's inequality:

$$|\langle x, y \rangle| \leq \|x\|_{p,p'} \|y\|_{q,q'}. \tag{3.3}$$

In particular,

$$|\langle x, y \rangle| \leq \|x\|_{1,2} \|y\|_{\infty,2}.$$

The direct consequence of Equation (3.3) is that the dual norm of an $\ell_{p,q}$ mixed norm is the $\ell_{p',q'}$ mixed norm, with $p'$ and $q'$ the respective Hölder conjugates of $p$ and $q$. Dual norms are of particular interest since they are closely linked to the definition of the subdifferential of our optimization problem.

The analytic notion of duality is analog to the geometric notion of polarity, which is useful to get a geometric insight into dual norms. A convex body $C^\star$ is a polar of another convex body $C$ if every point on its boundary defines a supporting hyperplane for $C$:

$$C^\star = \{x \in \mathbb{R}^p | \langle x, z \rangle \leq 0, \forall z \in C\}.$$

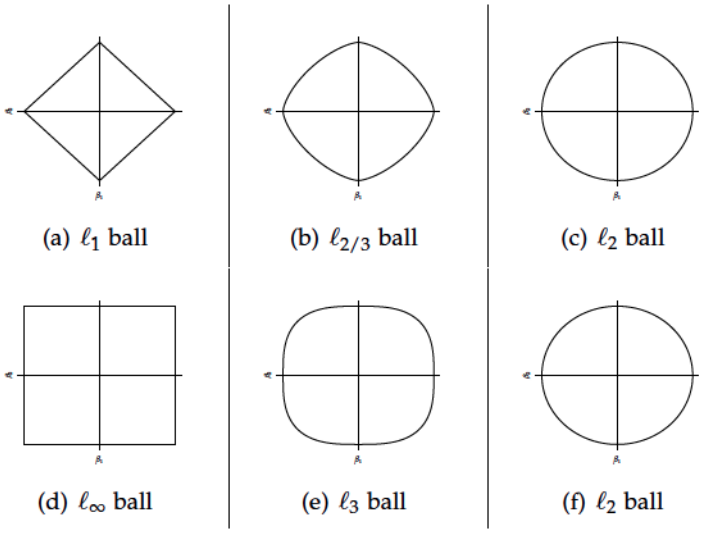
Duality and polarity are linked together by the fact that two norms are duals of each other if and only if their unit balls are polars of each other.

Some special cases prove straightforwardly the duality of familiar pair of norms. In particular, the polar of a sphere of radius $r$ is a sphere of radius $1/r$. For instance, the $\ell_2$ norm is its self-dual. Also, the polar of an intersection of closed half-spaces also shares a simple characterization. If there exists $(a_1, \ldots, a_H) \in \mathbb{R}^p$ such that $C = \{x \in \mathbb{R}^p, x^\intercal a_i \leq 1, \forall i = 1, \ldots, H\}$, then the polar is $C^\star = \text{conv}(a_1, \ldots, a_H)$, where conv denotes the convex hull. This result typically proves the duality between $\ell_1$ and $\ell_\infty$ norms.

As an illustration, Figure 3.1 provides the unit balls of three pairs of dual norms: $\ell_1 - \ell_\infty$, $\ell_{2/3} - \ell_3$ and $\ell_2 - \ell_2$.

### 3.1.2 Cooperative-Lasso Penalties as Sign-Adaptive Mixed Norms

As $\ell_p$ norms and mixed $\ell_{p,q}$ norms, let us define a set of cooperative $\ell_{p,q}$ norms as a generalization of the cooperative norm introduced in Chiquet et al. (2011). We prove in this Section that all the toolbox available form $\ell_p$ and mixed $\ell_{p,q}$ norms extend to cooperative norms.



(a) $\ell_1$ ball (b) $\ell_{2/3}$ ball (c) $\ell_2$ ball

(d) $\ell_\infty$ ball (e) $\ell_3$ ball (f) $\ell_2$ ball

Figure 3.1 – *Unit balls in $\mathbb{R}^2$ of three pairs of dual norms: $\ell_1$ and $\ell_\infty$ norms, $\ell_{2/3}$ and $\ell_3$ norms, $\ell_2$ and $\ell_2$ norm. For each norm $\|\cdot\|_p$ the boundary of the set $\{\beta \in \mathbb{R}^2, \|\beta\|_p \leq 1\}$ is drawn.*

**Definition 3.1** (Cooperative norms) *Consider $p, q \in [1, \infty]$. For every $x \in \mathbb{R}^p$ associated with a group structure $\{\mathcal{G}_k\}_{k=1}^K$, the $\ell_{p,q}$ coop-norm of $x$ is defined as a sign-adaptive mixed $\ell_{p,q}$ norm by:*

$$\|x\|_{coop,p,q} = \left( \sum_{k=1}^{K} \|x^+\|_q^p + \|x^-\|_q^p \right)^{1/p}.$$

*In particular,*

$$\begin{array}{lcll} \|x\|_{coop,1,2} & = & \|x^+\|_{1,2} + \|x^-\|_{1,2} & = \|x\|_{coop} \\ \|x\|_{coop,2,2} & = & \sqrt{\sum_{k=1}^K \|x^+\|_2^2 + \|x^-\|_2^2} & = \|x\|_2, \\ \|x\|_{coop,\infty,2} & = & \max\left(\|x^+\|_{\infty,2}, \|x^-\|_{\infty,2}\right) & = \|x\|_{coop^\star} \end{array}$$

.

It stems from this definition that the cooperative norm plays the role of a sign-adaptive mixed norm, the $\ell_q$ norm being taken on signed subgroups as if they were independent from each other.

Similarly to mixed norms, equivalence relationships generalize to cooperative norms, except that there is a factor 2 to be paid for the sign-adpativity in the compatibility constant.

**Proposition 3.1** (Equivalence relationships for coop-norms) *For every $x \in \mathbb{R}^p$,*

$$\begin{array}{lcll} \|x\|_2 & \leq & \|x\|_{coop} & \leq \sqrt{2K}\ \|x\|_2 \\ \|x\|_{coop,\star} & \leq & \|x\|_{coop} & \leq 2K\ \|x\|_{coop^\star} \\ \|x\|_{coop^\star} & \leq & \|x\|_2 & \leq \sqrt{2K}\ \|x\|_{coop^\star} \end{array}$$

Proposition 3.2 states the existence of dual bounds associated with cooperative norms and identifies the dual norm associated to $\|\cdot\|_{\text{coop}}$.

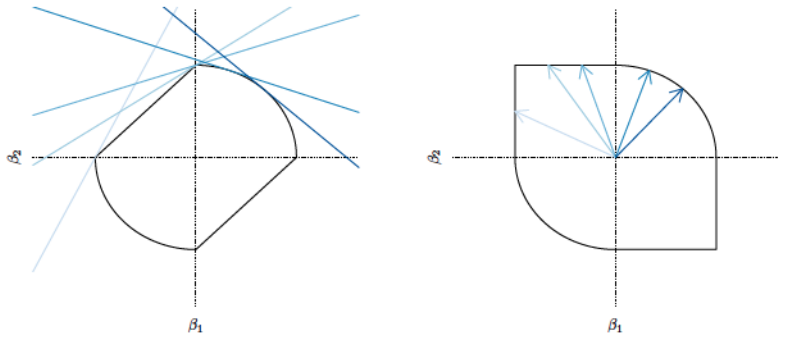
**Proposition 3.2** (Dual norm and dual bound for the coop-norm) *The dual norm of the coop-norm $\|\cdot\|_{coop} = \|\cdot\|_{coop,1,2}$ is $\|\cdot\|_{coop^\star} = \|\cdot\|_{coop,\infty,2}$, so that for every pair $(x, y) \in \mathbb{R}^2$, associated with a group structure $\{\mathcal{G}_k\}_{k=1}^K$, the following dual bound holds:*

$$|\langle x, y \rangle| \leq \|x\|_{coop} \|y\|_{coop^\star}.$$

*In general, for every two conjugate pairs $(p, q)$, $(p', q')$, the coop-$\ell_{p,q}$ norm is the dual of the coop-$\ell_{p',q'}$ norm, and they satisfy:*

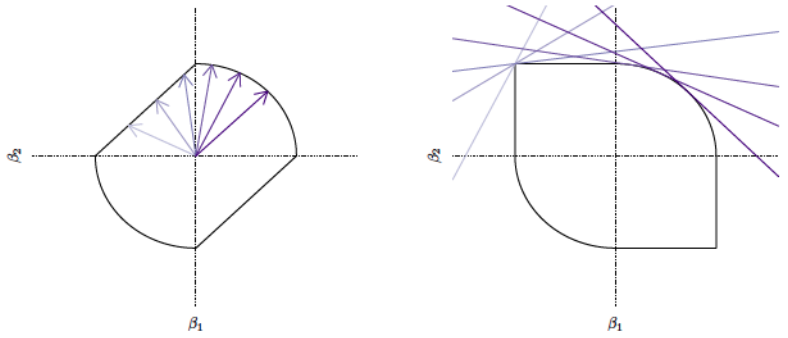
$$|\langle x, y \rangle| \leq \|x\|_{coop,p,p'} \|y\|_{coop,q,q'}.$$

The complete proof is postponed to Appendix A.1.1, but Figure 3.2 illustrates this duality in 2D, assuming $\beta_1$ and $\beta_2$ belong to one single group. On the left hand side appears the unit ball of the cooperative-norm, on the right hand side the unit ball of its dual. As illustrated, vectors of each unit ball define supporting hyperplanes to the dual norm ball.




(a) Coop ball and example of supporting hyperplanes defined by the Coop* ball

(b) Coop* ball and example of normal vectors to supporting hyperplanes of the Coop ball




(c) Coop ball and example of normal vectors to supporting hyperplanes of the Coop* ball

(d) Coop* ball and example of supporting hyperplanes defined by the Coop ball

Figure 3.2 – *The coop and coop* balls are polars of each other.*

## 3.2 The Cooperative-Lasso Problem and its Dual

Now that available tools are clarified, let us turn to the coop-Lasso optimization problem:

$$\hat{\boldsymbol{\beta}}^{\text{coop}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 + \lambda_n \|\boldsymbol{\beta}\|_{\text{coop}}. \tag{3.4}$$

Since the regularization is a norm, the problem is convex, which allows the derivation of various efficient optimization algorithms. However because

of the singularities in the cooperative-norm, the objective function is not differentiable. Wherever the gradient is not defined we need to resort to subgradients.

In this section, we start by recalling the definition of subgradients and subdifferentials in order to express the first-order optimality conditions of Problem 3.4. This leads us to the analysis of sparsity patterns achievable by the coop-Lasso. Then, we recall the definition of Fenchel conjugates in order to derive an analytical expression of first-order optimality conditions in terms of dual cooperative-norm. We conclude on the formulation of a dual problem associated to Problem 3.4. Both the expression of the subdifferential and the dual problem will be at the core of the proof for model selection.

### 3.2.1 Subdifferential and Achievable Sparsity Patterns

**Subdifferential and Subgradients.** Where the gradient provides a linear approximation to the objective function, Definition 3.2 recalls that the subgradient provides a lower-bound approximation.

**Definition 3.2** (Subgradients and subdifferential) *A vector $\theta \in \mathbb{R}^p$ is a subgradient to function $f : \mathbb{R}^p \to (-\infty, +\infty]$ at $x_0 \in \mathbb{R}^p$ if and only if for every $x \in \mathbb{R}^p$, $(\theta, -1)$ defines a supporting hyperplane from below to the curve representing $f$ at $x_0$:*

$$f(x) \geq f(x_0) + \langle \theta, x - x_0 \rangle.$$

*The subdifferential $\partial f(x_0)$ of $f$ at $x_0$ is the set of all subgradients of $f$ at $x_0$.*

Recall that the epigraph of a function $f$ is defined by $\text{epi}(f) = \{(x, y) \in \mathbb{R}^{p+1}, y \geq f(x)\}$ and that a normal cone to a set $C$ at $x_0$ is the convex cone such that $N_C(x_0) = \{y \in \mathbb{R}^p, < y, x - x_0 > \leq 0, \forall x \in C\}$. In geometric terms, Definition 3.2 states that when $f$ is convex, its subdifferential is linked to the normal cone of its epigraph, as illustrated by Figure 3.3.

**First-Order Optimality Conditions.** Naturally, as the notion of subgradient extends the notion of gradient, the subdifferential can be used to characterize the optimum of a function. Yet, as expressed in Proposition 3.3 and contrary to the gradient, the lower-bound approximation operated by the subdifferential does not characterize all (potentially local) optima, but only global minima.

**Proposition 3.3** (First-order condition : characterization of the optimum via the subdifferential) *The vector $x^\star$ is a global minimum to function $f : \mathbb{R}^p \to (-\infty, +\infty]$ with non-empty domain if and only if $0$ belongs to its subdifferential at $x^\star$:*

$$0 \in \partial f(x^\star).$$

When $f$ is differentiable, the subdifferential reduces to the derivative, and Proposition 3.3 boils down to the usual first-order condition $\nabla f(x^\star) = 0$. Therefore in the convex case, the subdifferential is only of interest when the function to minimize shows some singularities.

In our case, Problem 3.4 is a combination of a convex and differentiable function $\ell(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n / 2$ and a convex but non differentiable

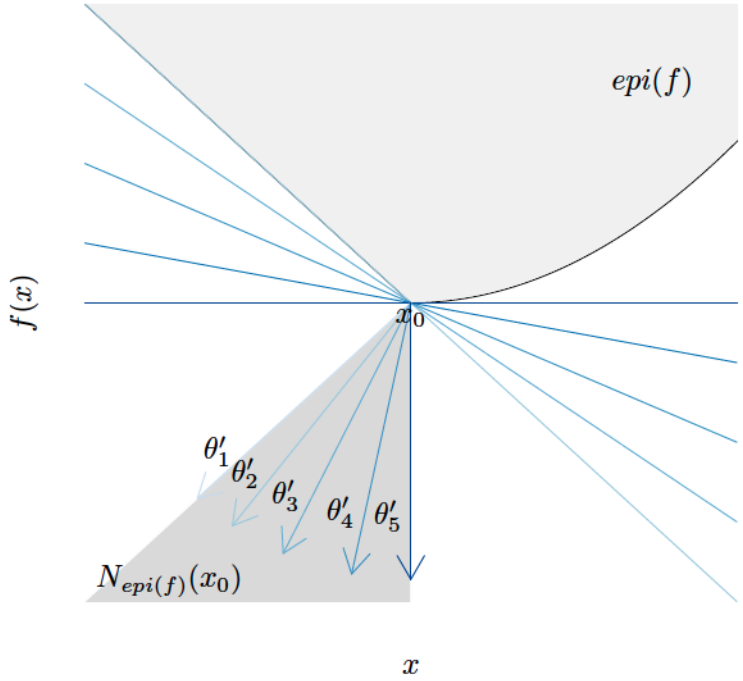



Figure 3.3 – *Subgradients $\theta_1, \dots \theta_5$ to $f$ at $x_0$ satisfy the condition that vectors $\theta'_1 = (\theta_1, -1), \dots, \theta'_5 = (\theta_5, -1)$ belong to the normal cone $N_{epi(f)}(x_0)$ to the epigraph of $f$ at $x_0$, $epi(f) = \{(x, y) \in \mathbb{R}^{p+1}, y \geq f(x)\}$.*

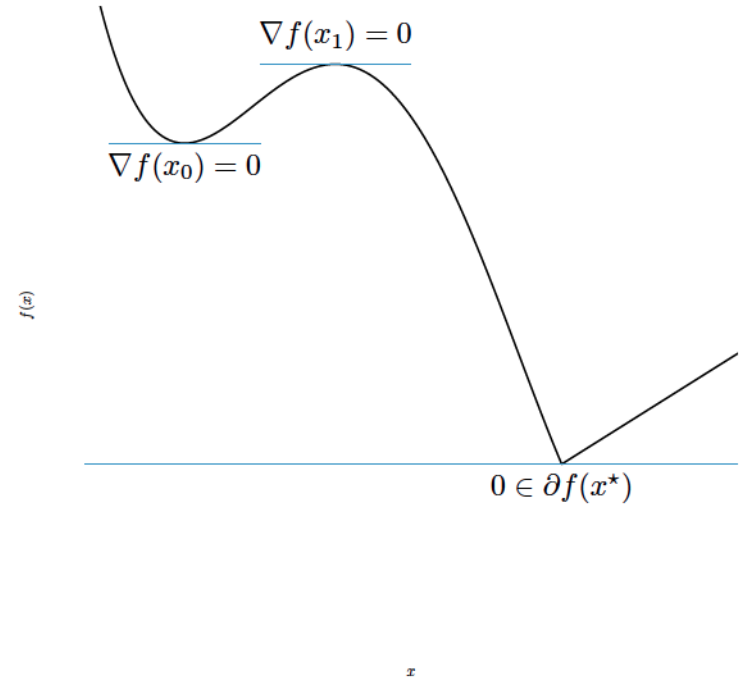



Figure 3.4 – *When the function $f$ is non-convex, the derivative charactizes local optima, while the subdifferential characterizes global minima.*

norm $\|\boldsymbol{\beta}\|_{\text{coop}}$. The vector $\hat{\boldsymbol{\beta}}$ is a global minimum to function $\ell(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) + \lambda f(\boldsymbol{\beta})$ if and only if $-\nabla\ell(\hat{\boldsymbol{\beta}}; \mathbf{X}, \mathbf{y})$ belongs to the subdifferential of $\|\boldsymbol{\beta}\|_{\text{coop}}$ at $\hat{\boldsymbol{\beta}}$:

$$-\nabla\ell(\hat{\boldsymbol{\beta}}; \mathbf{X}, \mathbf{y}) \in \partial\|\hat{\boldsymbol{\beta}}\|_{\text{coop}}. \tag{3.5}$$

Contrary to points $\hat{\boldsymbol{\beta}}$ where $\|\hat{\boldsymbol{\beta}}\|_{\text{coop}}$ is differentiable and its subdifferential $\partial\|\hat{\boldsymbol{\beta}}\|_{\text{coop}}$ reduces to the unique derivative $\nabla\|\hat{\boldsymbol{\beta}}\|_{\text{coop}}$, points with singularities have a non-zero probability to be selected as optimal, since a convex set of possible score vectors $\nabla\ell(\hat{\boldsymbol{\beta}}; \mathbf{X}, \mathbf{y})$ satisfy optimality conditions with respect to the same optimal point $\hat{\boldsymbol{\beta}}$. If singularities are placed at particular points of interest, admitting a specific support for instance, there is an increased probability for this particular support to be selected as optimal.

An interesting way of illustrating this phenomenon is to rephrase Problem 3.4 in terms of constrained least squares and amend Equation (3.5) accordingly. Instead of solving (3.4), minimize the sum of squared residuals under the constraint that $\|\boldsymbol{\beta}\|_{\text{coop}}$ remains smaller than $t$, $t > 0$. Under this formulation, a vector $\hat{\boldsymbol{\beta}}$ is optimum if and only if

$$-\nabla\ell(\hat{\boldsymbol{\beta}}; \mathbf{X}, \mathbf{y}) \in N_C(\hat{\boldsymbol{\beta}}), \tag{3.6}$$

that is to say, the score vector needs to belong to the normal cone to the feasible set $\mathcal{B}_{\text{coop}} = \{\boldsymbol{\beta} \in \mathbb{R}^p, \|\boldsymbol{\beta}\|_{\text{coop}} \leq t\}$ at the optimum. Normal cones to the coop ball of radius $t$ at singularities are represented on Figure 3.5.

Note that the normal cone to the feasible set $\mathcal{B}_{\text{coop}}$ at $x_0$ is nothing else than the polar of the set $\mathcal{B}_{\text{coop}} - x_0$, which links optimality conditions to the duality and polarity considerations of the previous section. The subdifferential associated with the coop-norm is the polar of the coop-norm unit ball, in other words the unit ball of the coop-dual norm. This point will be made clearer in the next subsection.

**Sparsity Patterns.** The constrained minimization formulation and Equation (3.6) implies that the solution to Problem 3.4 corresponds to the orthogonal projection of the ordinary least square estimate $\hat{\boldsymbol{\beta}}^{\text{ols}}$ onto a coop norm ball of a certain radius $t$. Coefficients will be set to 0 when level curves of the likelihood hit the ball at singularities, as illustrated in Figure 3.6.

Since the existence and position of singularities influence the range of amenable sparsity patterns, it is worth comparing the group, sparse group, and coop feasible sets. Consider for instance a vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^\intercal$ with two groups $\mathcal{G}_1 = \{1, 2\}$ and $\mathcal{G}_2 = \{3, 4\}$. Figure 3.7 presents cute 3D views of the unit balls according to $(\beta_1, \beta_2, \beta_3)$ for two different values of $\beta_4$, namely 0 and 0.3. To ease the identification of singularities, following figures provide various 2D cross-sections.

Figure 3.8 depicts within group cross-sections and illustrates the joint constraints imposed on $(\beta_1, \beta_2)$ for varying values of $\beta_4$ (0, and 0.3), $\beta_3$ being held at 0. Clearly in all three cases, groups are selected independently from each other: the activation of $\beta_4$ has no effect whatsoever on the singularities of within group balls, and therefore on the selection of either $\beta_1$ or $\beta_2$. Naturally though, since we are under the constrained formulation,

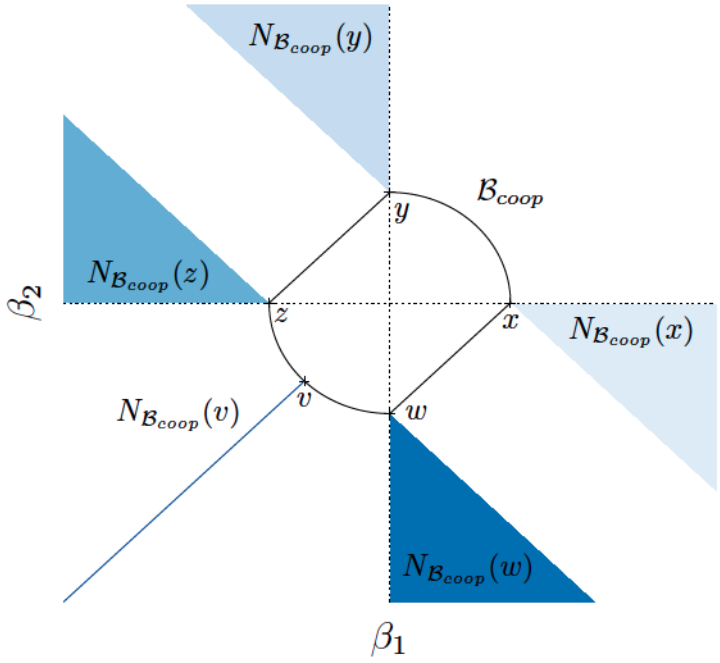



Figure 3.5 – *The coop norm ball of radius t in 2D $\mathcal{B}_{coop}$, on one group of size 2 $\{\beta_1, \beta_2\}$. Normal cones to $\mathcal{B}_{coop}$ at a differentiable point $v = (-sqrt(t/2), -sqrt(t/2))$ and specific points of singularities $w = (0, -t), x = (t, 0), y = (0, t), z = (-t, 0)$ are represented in blue. Contrary to the normal cones at $w, x, y, z$, the normal cone at $v$ is degenerated.*

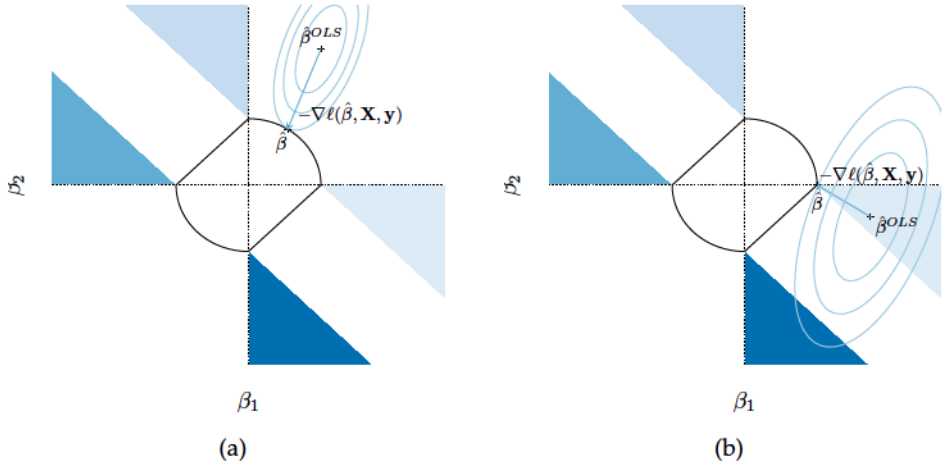



Figure 3.6 – *Projection $\hat{\boldsymbol{\beta}}$ of the Ordinary Least Square estimate $\hat{\boldsymbol{\beta}}^{\text{ols}}$ on the coop norm ball of radius t in 2D, with one group of size 2 $\{\beta_1, \beta_2\}$. On panel (a), projection hits on the $\mathbb{R}^+ \times \mathbb{R}^+$ quadrant: all variables are included. On panel (b), projection hits on the $\mathbb{R}^+ \times \mathbb{R}^-$ quadrant: $\hat{\boldsymbol{\beta}}_2$ can be set to 0.*

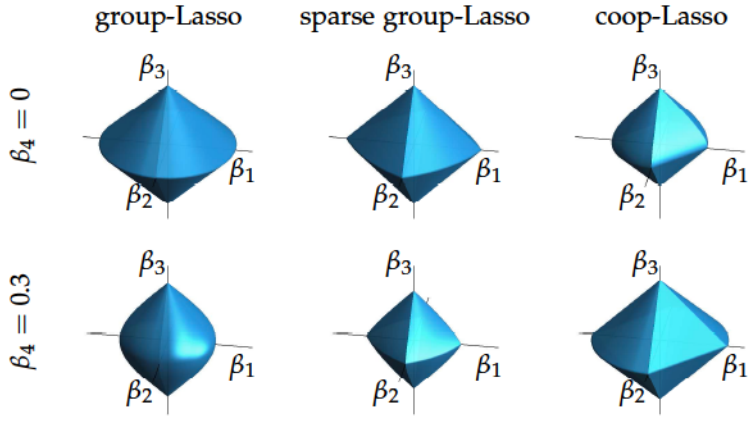



Figure 3.7 – *Feasible sets for the coop-Lasso, group-Lasso and sparse group-Lasso penalties. Cuts through $(\beta_1, \beta_2, \beta_3)$ at $\beta_4 = 0$ and $\beta_4 = 0.3$: $(\beta_1, \beta_2)$ span the horizontal plane and $\beta_3$ is on the vertical axis. These 3D views were realized by Yves Grandvalet.*

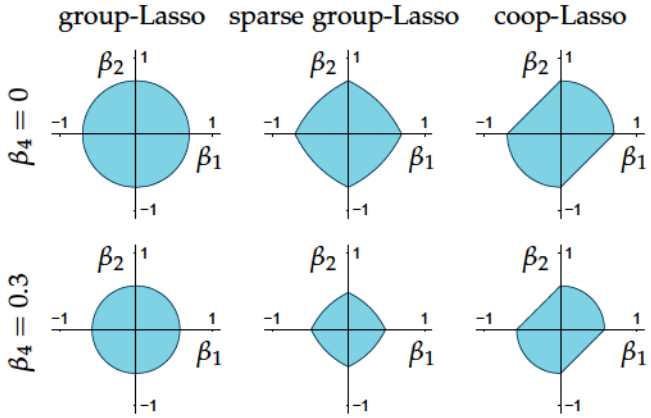
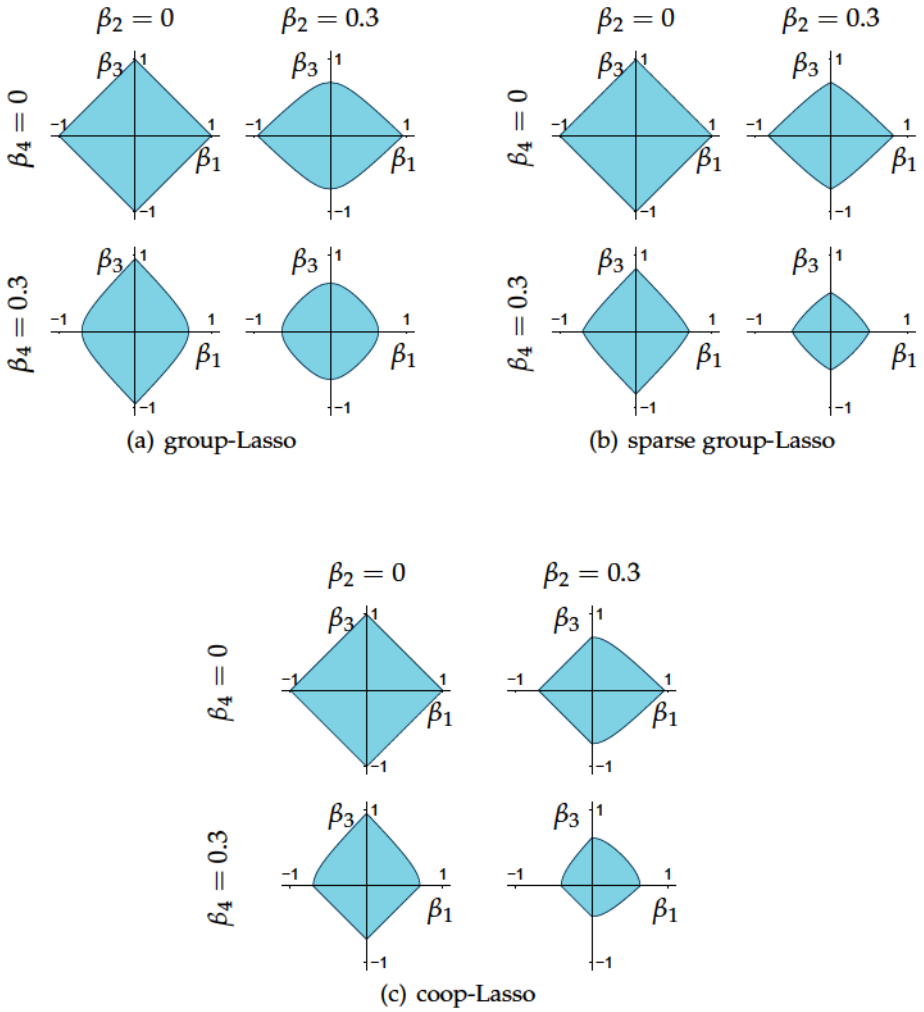



Figure 3.8 – *Feasible sets for the coop-Lasso, group-Lasso and sparse group-Lasso penalties. Cuts through $(\beta_1, \beta_2)$ at various values of $\beta_4$, and $\beta_3$ held fixed at 0.3. These 2D views were realized by Yves Grandvalet.*

the activation of coefficients in one group influences the size of coefficients in the other, hence the smaller radius of the 2D balls when $\beta_4 > 0$. As expected, the group feasible set presents no singularities, thereby illustrating the fact that the group Lasso cannot but activate all covariates within one group at the same time. Thanks to the $\ell_1$ supplementary regularization the sparse group feasible set presents singularities at all axes: the feasible set of the group-lasso gets shrunk towards the $\ell_1$ ball. The coop feasible set adds discontinuites on all axes but only on the sides of quadrants of diverging signs, as emphasized earlier by Figure 3.5.

Figure 3.9 focuses on across group cross-sections, representing $(\beta_1, \beta_3)$ for various values of $\beta_2$ and $\beta_4$. On the top-left panel, both $\beta_2$ and $\beta_4$ are switched off, while on the bottom-right panel both of them are activated. With all three types of norms, there naturally is a similar effect under the activation of either $\beta_2$ or $\beta_4$, with rotated but similar balls on top-right and bottom-left panels. With the group-Lasso, singularities disappear as soon as the other member of the group is activated. Singularities allowing to switch off $\beta_1$ (resp. $\beta_3$) disappear when $\beta_2$ (resp. $\beta_4$) is activated. As




(a) group-Lasso (b) sparse group-Lasso

(c) coop-Lasso

Figure 3.9 – *Feasible sets for the coop-Lasso, group-Lasso and sparse group-Lasso penalties. Cuts through* $(\beta_1, \beta_3)$ *at various values of* $(\beta_2, \beta_4)$*. These 2D views were realized by Yves Grandvalet.*

planned for, the sparse group-Lasso feasible sets maintain singularities at all corners, whatever be the value of other coefficients in the group, hence its ability to dissociate the selection of any variable from the selection of other variables in the same group. The coop-Lasso leads to more complex cuts. Here, singularities remain on the $\mathbb{R}^- \times \mathbb{R}^-$ quadrant in all panels since values considered for $\beta_2$ and $\beta_4$ are positive, but singularities on the positive side of the axis disappear when positive coefficients are activated in the group. Note that, in general, there are less new edges with the coop-lasso than with the sparse group-Lasso, since the new opportunities to switch off some coefficients are limited to the case where the group-Lasso would have allowed a solution with opposite signs within a group. The crucial difference between the coop- and the group- or sparse group-Lasso is the loss of the axial symmetry when some variables are non-zero: decoupling the positive and negative parts of the regression coefficients favors solutions where signs match within a group.

### 3.2.2 Fenchel Conjugate Functions and the Coop-Lasso Subdifferential

We now turn to the analytical expression of the coop-Lasso subdifferential, which was first exhibited in Chiquet et al. (2011). In this section, we derive the subdifferential for Problem 3.4 with the help of Fenchel conjugate functions and exhibit a possible dual form for Problem 3.4 in order to shed light on following consistency proofs. This section has been enriched by the reading of Bach et al. (2012), Boyd and Vandenberghe (2006), Borwein and Lewis (2006).

**Fenchel Conjugation.** Fenchel conjugate functions are of great help when deriving optimality conditions and dual problems. As represented on Figure 3.10, they measure the supremum gap between a linear function and the function $f$ of interest. The Fenchel conjugate $f^\star : \mathbb{R}^p \to [-\infty, \infty]$ of a function $f : \mathbb{R}^p \to [-\infty, \infty]$ is defined by :

$$f^\star(z) := \sup_{x \in \mathbb{R}^p} [\langle z, x \rangle - f(x)].$$

The Fenchel conjugate is always convex.

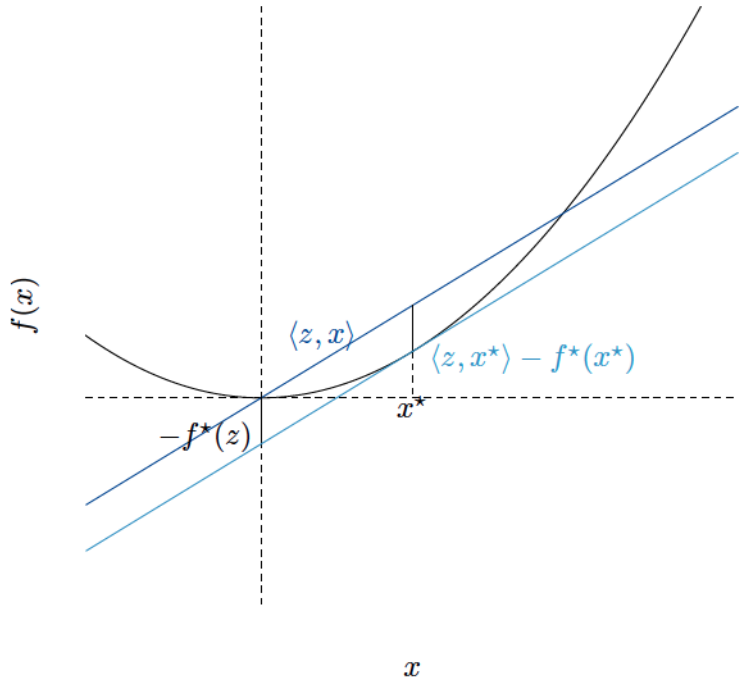



Figure 3.10 – *Construction of the Fenchel conjugate* $f^\star(z) := \sup_{x \in \mathbb{R}^p} [\langle z, x \rangle - f(x)] = \langle z, x^\star \rangle - f(x^\star)$. *When* $f$ *is differentiable,* $x^\star$ *is the point where the differential equals* $z$.

The Fenchel conjugate admits interesting expressions in the case of convex functions, norms in particular: the Fenchel conjugate of a norm $f : x \in \mathbb{R}^p \mapsto \|x\|$ is the indicator function of the unit ball of its dual norm $\|\cdot\|_\star$. For every $z$ in $\mathbb{R}^p$:

$$f^\star(z) = \begin{cases} 0 & \text{if } \|z\|_\star \leq 1 \\ +\infty & \text{otherwise} \end{cases}$$

**Analytic Expression of the Optimality Conditions.** The link between Fenchel conjugation and optimization is offered by Fenchel-Young inequality. The inequality has no real value in itself, since it is a straightforward consequence of the definition of the Fenchel conjugate. For every $x$ and $z$ in the domain of a function $f : \mathbb{R}^p \to (-\infty, +\infty]$, then:

$$f(x) + f^\star(z) \leq \langle x, z \rangle. \tag{3.7}$$

However, this inequality offers an efficient characterization of the subdifferential. Indeed, equality holds in (3.7) if and only if $z$ belongs to the subdifferential of $f$ at $x$, $\partial f(x)$. In particular, if $f$ is a norm $\|\cdot\|$, we obtain a concise expression for the subdifferential in terms of the associated dual norm $\|\cdot\|_\star$. For every $x$ and $y$ in $\mathbb{R}^p$,

$$z \in \partial f(x) \Leftrightarrow \begin{cases} \|z\|_\star \leq 1 & \text{if } x = 0 \\ \|z\|_\star = 1 \text{ and } \langle x, z \rangle = \|x\| & \text{otherwise.} \end{cases} \tag{3.8}$$

It suffices to combine this characterization of the subdifferential with Equation (3.5) and the definition of the dual coop norm to obtain explicit optimality conditions for the cooperative-Lasso. Since the coop-Lasso acts as a sign-adaptive group-Lasso, let us introduce some new notations to clarify the following results. For each group $\mathcal{G}_k$, define $s_{2k-1}$ and $s_{2k}$ the signed subsets of respectively positive and negative coefficients in $\hat{\boldsymbol{\beta}}_{\mathcal{G}_k}$.

**Theorem 3.4** (Optimality conditions for the cooperative-Lasso) *The vector $\hat{\boldsymbol{\beta}}$ is optimal for Problem 3.4 if and only if $z = -\mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/w_k\lambda$ belongs to the subdifferential of the coop-norm associated to groups $\{\mathcal{G}_k\}_{k=1}^K$ at $\hat{\boldsymbol{\beta}}$, characterized by indicator function of the coop-dual norm $\|\cdot\|_{coop^\star}$. The vector $\hat{\boldsymbol{\beta}}$ is optimal if and only if, for every group $\mathcal{G}_k$,*

$$\max(\|z^+_{\mathcal{G}_k}\|_2, \|z^-_{\mathcal{G}_k}\|_2) \leq 1,$$

*and, in particular, for every group $\mathcal{G}_k$ such that positive and/or negative coefficients are activated:*

$$z_{s_j} = \hat{\boldsymbol{\beta}}_{s_j}\|\hat{\boldsymbol{\beta}}_{s_j}\|^{-1} \quad \text{for } j = 2k-1 \text{ or } 2k, \text{ such that } s_j \neq \emptyset.$$

A strong consequence of Theorem 3.4 is that if both positive and negative coefficients are activated within the same group, then no other coefficients can be shrunk to zero in that group.

### 3.2.3 The Dual Problem

We devote a small subsection to the formulation of a dual problem associated with the coop-Lasso derived thanks to Fenchel conjugation. Dual problems can be used in optimization algorithms in order to check the convergence of the algorithm. The main motivation here for the formulation of this problem is the primal-dual witness construction of support recovery results.

By definition of the cooperative dual norm, Problem 3.4 can be rewritten under the following primal form:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 + \lambda \, \|\boldsymbol{\beta}\|_{\text{coop}} \Leftrightarrow \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sup_{\|\boldsymbol{\alpha}\|_{\text{coop}^\star} \leq \lambda_n} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 - \langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle.$$

This primal problem admits the dual formulation given in Proposition 3.5.

**Proposition 3.5** (Dual problem) *Problem 3.4 admits the following dual problem:*

$$\sup_{\|\boldsymbol{\alpha}\|_{coop^\star} \leq \lambda_n} -f^\star(\boldsymbol{\alpha}).$$

*where $f^\star$ is the Fenchel conjugate of $f : \boldsymbol{\beta} \mapsto f(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2/2$.*

*Proof.* Applying the *min* − *max* inequality and following the definitions of the cooperative dual norm and Fenchel conjugate of

$$\begin{aligned}
\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sup_{\|\boldsymbol{\alpha}\|_{\text{coop}^\star} \leq \lambda_n} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 - \langle \boldsymbol{\alpha}, \boldsymbol{\beta}\rangle &\geq \sup_{\|\boldsymbol{\alpha}\|_{\text{coop}^\star} \leq \lambda_n} \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 - \langle \boldsymbol{\alpha}, \boldsymbol{\beta}\rangle \\
&\geq \sup_{\|\boldsymbol{\alpha}\|_{\text{coop}^\star} \leq \lambda_n} - \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \langle \boldsymbol{\alpha}, \boldsymbol{\beta}\rangle - \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 \\
&\geq \sup_{\|\boldsymbol{\alpha}\|_{\text{coop}^\star} \leq \lambda_n} -f^\star(\boldsymbol{\alpha}).
\end{aligned}$$

□

Note that $f^\star(\boldsymbol{\alpha})$ is obtained at points $\boldsymbol{\beta}$ such that

$$\nabla f(\boldsymbol{\beta}) = \mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\alpha}.$$

Moreover, both terms of the primal-dual inequality are equal as soon as the domain of $f$ has non-empty interior, and in that case the duality gap $f(\hat{\boldsymbol{\beta}}) + \langle \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\rangle - f^\star(\hat{\boldsymbol{\alpha}})$ between the left and right terms of the inequality reduces to 0. In that condition, we say that strong duality holds. As a result of Fenchel-Young inequality, it appears that when strong duality indeed holds, the dual optimal variable $\hat{\boldsymbol{\alpha}}$ belongs to the subdifferential at $\hat{\boldsymbol{\beta}}$. This observation is of upmost importance to understand the proof of model selection consistency.

## 3.3 Consistency

Beyond its sanity-check value, a consistency analysis brings along an appreciation of the strengths and limitations of an estimation scheme. We provide two types of results, based upon best achievable results for the Lasso. First, we derive selection properties in an asymptotic linear regression framework, based upon an irrepresentable condition which is the analogue of the sufficient and necessary condition for the selection consistency of the Lasso. Secondly, we prove estimation and prediction sparsity oracle inequalities, valid non-asymptotically and based upon a Restricted Eigenvalue assumption.

### 3.3.1 Asymptotic Properties as a Selection Tool

Here we concentrate on the estimation of the support of the parameter vector, that is, the position of its zero entries. Our proof technique is

drawn from the previous works on the Lasso (Yuan and Lin 2007b) and the group-Lasso (Bach 2008b).

In this type of analysis, some assumptions on the joint distribution of $(X, Y)$ are required to guarantee the convergence of empirical covariances. For the sake of simplicity and coherence, we keep assuming that data are centered so that we have zero mean random variables and $\mathbf{\Psi} = \mathbb{E}\left[XX^\intercal\right]$ is the covariance matrix of $X$.

(A1) $X$ and $Y$ have finite 4th order moments $\mathbb{E}\left[\|X\|^4\right] < \infty$, $\mathbb{E}\left[Y^4\right] < \infty$.

(A2) The covariance matrix $\mathbf{\Psi} = \mathbb{E}\left[XX^\intercal\right] \in \mathbb{R}^{p\times p}$ is invertible.

In addition to these standard technical assumptions, we need a more specific one, substantially avoiding situations where the coop-Lasso will almost never recover the true support $\mathcal{S}$. In the sequel, $\mathcal{S}$ denotes the true support of $\boldsymbol{\beta}^\star$, while $\mathcal{S}_k$ denotes the intersection between the support and group $\mathcal{G}_k$.

(A3) All sign-incoherent groups are included in the true support: $\forall k \in \{1,\dots,K\}$, if $\|(\boldsymbol{\beta}^\star_{\mathcal{G}_k})^+\| > 0$ and $\|(\boldsymbol{\beta}^\star_{\mathcal{G}_k})^-\| > 0$, then $\forall j \in \mathcal{G}_k$, $\boldsymbol{\beta}^\star_j \neq 0$.

Note that this latter assumption is less stringent than the one required for the group-Lasso since it does not require that each group of variables should either be included in or excluded from the support. For the coop-Lasso, sign-coherent groups may intersect the support.

The spurious relationships that may arise from confounding variables are controlled by the so-called strong irrepresentable condition, which guarantees support recovery for the Lasso (Yuan and Lin 2007b) and the group-Lasso (Bach 2008b). We now introduce suitable variants of these conditions for the coop-Lasso. They result in two assumptions: a general one, on the magnitude of correlations between relevant and irrelevant variables, and a more specific one for groups which intersect the support, on the sign of correlations. These conditions will be expressed in a compact vectorial form using the diagonal weighting matrix $\mathbf{D}(\boldsymbol{\beta})$ such that,

$$\forall k \in \{1,\dots,K\},\ \forall j \in \mathcal{S}_k(\boldsymbol{\beta}),\ \ \left(\mathbf{D}(\boldsymbol{\beta})\right)_{jj} = w_k \|\boldsymbol{\varphi}_j(\boldsymbol{\beta}_{\mathcal{G}_k})\|^{-1}\ , \tag{3.9}$$

where $\boldsymbol{\varphi}_j(\boldsymbol{\beta}_{\mathcal{G}_k})$ is the restriction of $\boldsymbol{\beta}_{\mathcal{G}_k}$ to the subset of its components which share the same sign as $\boldsymbol{\beta}_j$.

(A4) For every group $\mathcal{G}_k$ including at least one null coefficient (that is, such that $\beta^\star_j = 0$ for some $j \in \mathcal{G}_k$ or equivalently $\mathcal{S}_k^c \neq \emptyset$), there exists $\eta > 0$ such that

$$\frac{1}{w_k}\|(\mathbf{\Psi}_{\mathcal{S}_k^c\mathcal{S}}\mathbf{\Psi}^{-1}_{\mathcal{S}\mathcal{S}}\mathbf{D}(\boldsymbol{\beta}^\star_{\mathcal{S}})\boldsymbol{\beta}^\star_{\mathcal{S}})\|_{\text{coop}^\star} \leq 1-\eta\ , \tag{3.10}$$

where $\mathbf{\Psi}_{\mathcal{S}\mathcal{T}}$ is the submatrix of $\mathbf{\Psi}$ with lines and columns respectively indexed by $\mathcal{S}$ and $\mathcal{T}$.

(A5) For every group $\mathcal{G}_k$ intersecting the support and including either positive or negative coefficients, let $\nu_k$ be the sign of these coefficients

($\nu_k = 1$ if $\|(\boldsymbol{\beta}^\star_{\mathcal{G}_k})^+\| > 0$ and $\nu_k = -1$ if $\|(\boldsymbol{\beta}^\star_{\mathcal{G}_k})^-\| > 0$), the following inequalities should hold:

$$\nu_k \boldsymbol{\Psi}_{\mathcal{S}_k^c \mathcal{S}} \boldsymbol{\Psi}_{\mathcal{S}\mathcal{S}}^{-1} \mathbf{D}(\boldsymbol{\beta}^\star_{\mathcal{S}}) \boldsymbol{\beta}^\star_{\mathcal{S}} \preceq \mathbf{0} \ , \tag{3.11}$$

where $\preceq$ denotes componentwise inequality.

Note that the irrepresentable condition for the group-Lasso only considers correlations between groups included and excluded from the support. It is otherwise similar to (3.10), except that the elements of the weighting matrix $\mathbf{D}$ are $w_k \|\boldsymbol{\beta}_{\mathcal{G}_k}\|^{-1}$ and that the self dual $\ell_2$ norm replaces the dual cooperative norm.

**Theorem 3.6** *If assumptions (A1-5) are satisfied, the coop-Lasso estimator is asymptotically unbiased and has the property of exact support recovery:*

$$\hat{\boldsymbol{\beta}}_n^{\text{coop}} \xrightarrow{P} \boldsymbol{\beta}^\star \quad \textit{and} \quad \mathbb{P}\left(\mathcal{S}(\hat{\boldsymbol{\beta}}_n^{\text{coop}}) = \mathcal{S}\right) \to 1 \ , \tag{3.12}$$

*for every sequence $\lambda_n$ such that $\lambda_n = \lambda_0 n^{-\gamma}$, $\gamma \in (0, 1/2)$.*

Compared to the group-Lasso, the consistency of support recovery for the coop-Lasso differs primarily regarding possible intersection (besides inclusion and exclusion) between groups and support. This additional flexibility applies to every sign-coherent group. Even if the support is the union of groups, when all groups are sign-coherent, the coop-Lasso has still an edge on group-Lasso since the irrepresentable condition (3.10) is weaker. Indeed, the norm in (3.10) is dominated by the $\ell_2$ norm used for the group-Lasso. The next paragraph illustrates that this difference can have remarkable outcomes. Finally, when the support is the union of groups comprising sign-incoherent ones, there is no systematic advantage in favor of one or the other method. While the norm used by the coop-Lasso is dominated by the norm used by the group-Lasso, the weighting matrix $\mathbf{D}$ has smaller entries for the latter.

**Remark 3.1** *Contrary to Negahban and Wainwright (2011) and Obozinski et al. (2011), those results do not apply to high-dimensional settings where the number of variables exceeds the sample size. In order to adapt Theorem 3.6, one would need to add technical assumptions guaranteeing the existence of concentration inequalities, however assumptions required on the design to obtain exact support recovery would remain the same as in (A4) and (A5). Since the latter are the only assumptions that would differ between Obozinski et al. (2011) and the coop-Lasso, there is no major interest in rewriting Theorem 3.6 according to those new developments except for the pleasure of reading them in their up-to-date formulation.*

**Illustration** We generate data from an ordinary regression model with $\boldsymbol{\beta}^\star = (1, 1, -1, -1, 0, 0, 0, 0)$, equipped with the group structure $\{\mathcal{G}_k\}_{k=1}^4 = \{\{1,2\}, \{3,4\}, \{5,6\}, \{7,8\}\}$. The vector $X$ is generated as a centered Gaussian random vector whose covariance matrix $\boldsymbol{\Psi}$ is chosen so that the irrepresentable conditions hold for the coop-Lasso, but not for group-Lasso, which, we recall, are more demanding for the current situation, with sign-coherent groups. The random error $\varepsilon$ follows a centered Gaussian distribution with standard deviation $\sigma = 0.1$, inducing a

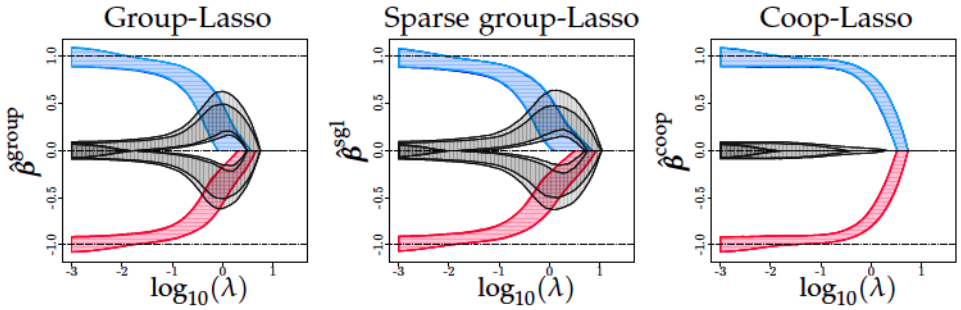



Figure 3.11 – *50% coverage intervals for the group (left), sparse group (center), and (right) Lasso estimated coefficients along regularization paths: coefficients from the support of $\beta^\star$ are marked by colored horizontal stripes and the other ones by gray vertical stripes.*

very high signal to noise ratio ($R^2 = 0.99$ on average), so that asymptotics provide a realistic view of the finite sample situation.

We generated 1000 samples of size $n = 20$ from the described model, computed the corresponding 1000 regularization paths for the group-Lasso, sparse group-Lasso, and coop-Lasso. Figure 3.11 reports the 50% coverage intervals (lower and upper quartiles) along the regularization paths. In this setup, the sparse group-Lasso behaves as the group-Lasso, leading to nearly identical graphs. Estimation is difficult in this small sample problem ($n = 20, p = 8$), and the two versions of the group-Lasso, which first select the wrong covariates, never reach the situation where they would have a decisive advantage upon OLS, while the coop-Lasso immediately selects the right covariates, whose coefficients steadily dominate the irrelevant ones. Model selection is also difficult, and the BIC criteria provided in Section 1.3.1 select often the OLS model (in about 10% and 50% of cases for the coop-Lasso and the group-Lasso respectively). The average root mean square error on parameters is of order $10^{-1}$ for all methods, with a slight edge for coop-Lasso. The sign error is much more contrasted: 31% for the coop-Lasso *vs.* 46% for the group-Lasso, not far better than the 50% of OLS.

### 3.3.2 Non-Asymptotic Properties for Estimation and Prediction Purposes

In the high-dimensional framework, where the number of observations is small compared to the number of variables, it is crucial to understand the non-asymptotic properties of the estimator. In that respect, we derive non-asymptotic oracle inequalities, based upon restricted eigenvalue assumptions.

Similarly to the Lasso, bounds on estimation and prediction error for the cooperative-Lasso are subject to restricted strong convexity assumptions. The assumption is roughly speaking the same, except that the cone on which the assumption relies is defined by the cooperative-norm, and the sparsity considered is a group sparsity.

**Assumption 3.1** (Restricted eigenvalue) *There exists $\kappa(s) > 0$ such that:*

$$\min \left\{ \frac{\|Xu\|_2}{\sqrt{n}\|u_{\mathcal{S}}\|_2} : |\mathcal{S}| \leq s, u \in \mathbb{R}^p, \|u_{\mathcal{S}^c}\|_{coop} \leq 3\|u_{\mathcal{S}}\|_{coop} \right\} \geq \kappa(s)$$

Under assumption 3.1 and for a good choice of the tuning parameter $\lambda_n$, we obtain prediction and estimation bounds as in Theorem 3.7.

**Theorem 3.7** (Oracle inequalities) *Under Assumption 3.1, for a choice of $\lambda_n \geq 2\|\mathbf{X}^\intercal \varepsilon/n\|$, any solution to Problem 3.4 satisfies the following prediction and estimation oracle inequalities:*

$$\|X(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n^2 \leq \frac{32\lambda_n^2}{\kappa(s)^2} s,$$
$$\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{coop} \leq \frac{32\lambda_n}{\kappa(s)^2} s,$$
$$\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_2 \leq \frac{32\lambda_n}{\kappa(s)^2} s.$$

In order to explicit the bounds and order of probability in Theorem 3.7, let us restrict ourselves to the case where all groups share the same size $m$.

**Corollary 3.8** (Oracle inequalities with groups of equal sizes) *Under Assumption 3.1 and considering that the data matrix has been scaled so that all diagonal elements of $\mathbf{X}^\intercal\mathbf{X}/n$ are equal to 1, for a choice of $\lambda_n$ equal to $\sigma\frac{\sqrt{m}+\sqrt{\log K}}{\sqrt{n}}$, then with probability larger than $1 - 2/K^2$, any solution to Problem 3.4 with groups of equal sizes $m$ satisfies the following prediction and estimation oracle inequalities:*

$$\|X(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n^2 \leq \frac{32}{\kappa(s)^2} \frac{s(m + \log K + 2\sqrt{m \log K})}{n}, \tag{3.13}$$
$$\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{coop} \leq \frac{32}{\kappa(s)^2} \frac{s(\sqrt{m} + \sqrt{\log K})}{\sqrt{n}}, \tag{3.14}$$
$$\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_2 \leq \frac{32}{\kappa(s)^2} \frac{s(\sqrt{m} + \sqrt{\log K})}{\sqrt{n}}. \tag{3.15}$$

**Remark 3.2** *We need to restrict ourselves to groups of equal size because the upper bound on the probability of the event $\{\lambda_n \geq 2\|\mathbf{X}^\intercal\varepsilon/n\|\}$ for fixed $\lambda_n$ relies on tail bounds of the maximum of K chi-square distributions. If all groups share the same size, then we can easily use a union bound on the tails of K independent chi-square with similar degrees of freedom. Otherwise, each of the K chi-square distributions has its own degree of freedom, which makes it impossible to upper-bound explicitely the probalitity of the intersection, unless we use a very raw upper-bound.*

In the case of multitask data, where the number of groups actually corresponds to the number of variables $p$, the group size is equal to the number of tasks $T$, and the number of observations $n$ is in fact equal to $NT$ where $N$ is the number of observations gathered by condition, the $\ell_2$ bound reads:

$$\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_2 \leq \frac{32}{\kappa(s)^2} \frac{s}{\sqrt{N}} \left(1 + \sqrt{\frac{\log p}{T}}\right).$$

The rate $\frac{s}{\sqrt{N}}$ shows that like the group-Lasso, the coop-Lasso adapts to the unknown group-sparsity, but not without paying a price of the order

$\sqrt{\log p}$ for not knowing in advance the truely relevant groups. The $T$ task replicates alleviate this cost by a factor $\sqrt{T}$.

**Remark 3.3** *There is no improvement compared to the group-Lasso oracle inequalities because we cannot exploit the advantages of the cooperative-norm on two fronts. First, the probability of event $\mathcal{A}$ uses an upper-bound of the dual cooperative-norm by the dual group-norm, because the dual coop-norm leads to chi-square distributions of unknown degrees of freedom which we cannot control explicitly. Second, what appears is actually a rate of $2s$, twice the group-sparsity: we cannot count the number of activated signed-groups instead. Indeed, following the terms of S. Negahban and Yu (2012) the cooperative-norm is only decomposable to group-sparse subsets, not to signed quadrants: we can write $\|\alpha+\beta\|_{coop} = \|\alpha\|_{coop} + \|\beta\|_{coop}$ for every $\alpha \in \mathcal{M}$ and $\beta \in \mathcal{M}^{\perp}$ for subsets $\mathcal{M} = \{x \in \mathbb{R}^p, \forall k \in S^c, x_{\mathcal{G}_k} = 0\}$ defined by the activation of a subset of groups, but not by the activation of a subset of* signed *subgroups.*

*Recent developments in S. Negahban and Yu (2012) and Lounici et al. (2011) could help improve the results. The former allows to consider weakly group-sparse vectors defined as $\ell_{2,q}$ bounded vectors instead of group sparse vectors. The latter allows to work with group-specific penalties $\lambda_j$, which might be interesting for at least two reasons: first, in order to adapt the amount of penalty to the size of the groups, second in order to derive oracle inequalities for weighted cooperative regularizations. Again, the main difference between the group-Lasso and coop-Lasso results lies in the Restricted Eigenvalue assumptions. As a result, there is no major novelty to be learnt from the adaptation of those recent results to the case of the coop-Lasso.*

## 3.4 Application to the Inference of Multiple Gaussian Graphical Models

As exposed in introduction, the first motivation of the cooperative-Lasso is the inference of multiple Gaussian graphical models as in Chiquet et al. (2011). This section details the adaptation of the cooperative-Lasso for linear regression to the inference of multiple Gaussian graphical models. This application is illustrated on two real datasets.

### 3.4.1 Statistical Modeling

Let us model expression levels by condition-specific Gaussian distributions with condition-specific means $\mu^{(c)}$ – which vanish when centering the data condition by condition – and covariance matrix $\Sigma^{(c)}$. In each condition, the distribution of the expression vector $X^{(c)}$ is modeled by a Gaussian graphical model with graph of conditional dependencies $\Gamma^{(c)}$, whose edges correspond to the non-zero entries of the inverse covariance matrix $\Theta^{(c)}$. For clarity reasons, we assume that we gather the same number $n$ of observations in each condition, stored in an $n \times p$ matrix $\mathbf{X}^{(c)}$. Under the natural assumption of independence between conditions, the log-likelihood within each sample $c$ admits the same form as in Problem 1.10 and the log-likelihood of the overall sample

$\mathbf{X} = (\mathbf{X}^{(1)} \ \dots \ \mathbf{X}^{(c)} \ \dots \ \mathbf{X}^{(C)}) \in \mathcal{M}_{n \times Cp}$ factorizes into:

$$\ell(\mathbf{X}; \mathbf{\Theta}) = \frac{n}{2} \left[ \sum_{c=1}^{C} \log\det \Theta^{(c)} - \langle \mathbf{S}^{(c)}, \Theta^{(c)} \rangle - Cp \log 2\pi \right]$$

where $\mathbf{\Theta} = (\Theta^{(1)} \ \dots \ \Theta^{(c)} \ \dots \ \Theta^{(C)})$ to alleviate notations and $\mathbf{S}^{(c)}$ denotes the empirical covariance matrix in condition $c$.

Similarly to the i.i.d. setting, this problem can be solved via an appropriate regularization of the likelihood:

$$\widehat{\mathbf{\Theta}} = \arg\min \mathcal{L}(\mathbf{X}; \mathbf{\Theta}) + \lambda \ \text{pen}(\mathbf{\Theta}). \tag{3.16}$$

Adopting a neighborhood selection strategy, regularizations presented in the previous sections in the general linear regression framework are directly transposable to the case of multiple Gaussian graphical models. Instead of solving Problem (3.16), $p$ independent problems like Problem (3.18), one for each gene $g$.

$$\min_{\boldsymbol{\beta}_g} \sum_{c=1}^{C} \|\mathbf{X}_g^{(c)} - \mathbf{X}_{\setminus g}^{(c)} \beta_g^{(c)}\|_n^2 + \lambda_n \text{pen}(\boldsymbol{\beta}_g). \tag{3.17}$$

Each Problem (3.18) can be rephrased as a unique linear regression with partionned variables as originally in Equation (3.1). Denote by $\mathbf{X}_{\setminus g}$ the $nC \times n(p-1)$ block diagonal matrix formed with the $\{\mathbf{X}_{\setminus g}^{(c)}\}, c = 1, \dots, C$, by $\mathbf{X}_g$ the size-$nC$ vector concatenating the observations for gene $g$ in the $C$ conditions, and finally by $\boldsymbol{\beta}_g$ the size $n(p-1)$ vector concatenating all $\beta_g^{(c)}$'s, for $c = 1, \dots, C$. Then Problem (3.18) is equivalent to

$$\min_{\boldsymbol{\beta}_g} \|\mathbf{X}_g - \mathbf{X}_{\setminus g} \boldsymbol{\beta}_g\|_n^2 + \lambda_n \text{pen}_{\mathcal{G}}(\boldsymbol{\beta}_g), \tag{3.18}$$

using the partition

$$\mathcal{G} = (\underbrace{1, 2, \dots, C}_{p-1 \text{ times}}, 1, 2, \dots, C, \dots).$$

The different regularization terms pen($\mathbf{\Theta}$) presented in the previous sections will result in the estimation of different sparsity patterns linked to different assumptions on the amount of heterogeneity across conditions. Indeed, we assume that measurements in all conditions focus on the activity of the same set of genes $\mathcal{P}$, but need to loosen the fundamental i.i.d. assumption across conditions in different ways that we detail now. We are focus on four different assumptions on the similarities between partial correlations structures across conditions, as four different condition-specific variations around a common structure represented by $\Sigma^\star$, $\Theta^\star$ and $\Gamma^\star$.

(C1) *Identically distributed:* All conditions share the same covariance and therefore concentration matrices $\Sigma^\star$ and $\Theta^\star$. For every condition $c$ and every pair of genes $(g, h) \in \mathcal{P}^2$

$$\theta_{gh}^{(c)} = \theta_{gh}^\star;$$

(C2) *Identical partial correlation structures*: All conditions share the same graph of conditional dependencies $\Gamma^\star$. For every condition $c$ and every pair of genes $(g,h) \in \mathcal{P}^2$,

$$\theta^{(c)}_{gh} \neq 0 \Leftrightarrow \theta^\star_{gh} \neq 0, \quad \text{i.e.} \quad \Gamma^{(c)} = \Gamma^\star;$$

(C3) *Almost identical partial correlation structures*: All conditions share the same graph of conditional dependencies $\Gamma$ except for a small set of condition specific edges, such that for every condition $c$ and every pair of genes $(g,h) \in \mathcal{P}^2$

$$\theta^{(c)}_{gh} \neq 0 \Rightarrow \theta^\star_{gh} \neq 0, \quad \text{i.e.} \quad \Gamma^\star \subseteq \Gamma^{(c)};$$

Note that this assumption encompasses cases where edges would be missing in some conditions when compared to the common benchmark $\Gamma^\star$. This configuration fit into assumption (C3) as long as $\Gamma^\star$ is defined as the reunion of all condition-specific graphs.

(C4) *Almost identical sign-coherent partial correlation structures*: All conditions share the same graph of conditional dependencies $\Gamma$, with all $\Theta^{(c)}$'s sharing the exact same sign-pattern, except that, for each edge, there can exist a subset of disagreeing conditions where this edge can either disappear or switch to the opposite sign. For every pair of genes $(g,h) \in \mathcal{P}^2$, the edge $(g,h)$ either shares the majority sign $\text{sign}(\theta^\star_{gh})$ or the minority sign $s_{gh}$. For every condition $c$ and pair of genes $(g,h) \in \mathcal{P}^2$,

$$\text{sign}(\theta^{(c)}_{gh}) \in \{\text{sign}(\theta^\star_{gh}), s_{gh\cdot}\}$$

Note that even though this assumption is called *sign-coherent*, it allows for positive edges in a majority of conditions to cohabit with negative edges in the remaining conditions. However it forbids situations where the same edge would be positive in some conditions, negative in others, and absent in a last subset of conditions.

Under any of those four assumptions, combining observations from different conditions into one single inference problem is a way of alleviating the burden of high-dimension, in the spirit of panel datasets or multi-task experiments. Assumption (C1) corresponds to the highly unlikely case where all conditions actually happen to form a larger i.i.d. dataset. We only add it to stress the fact that discarding the heterogeneity and naively merging all information into one single estimator is doomed to be incorrect.

Assumption (C2) alleviates assumption (C1) by allowing edges to differ in intensity across conditions. This setting corresponds to the group-Lasso penalty:

$$\text{pen}_{\text{group}}(\boldsymbol{\Theta}) = \sum_{(g,h)\in\mathcal{P}^2} \left( \sum_{c=1}^{C} (\theta^{(c)}_{gh})^2 \right)^{1/2}.$$

As assumption (C2) suggests, the group-Lasso results in different $\widehat{\Theta}^{(c)}$ estimates but in a single common conditional dependency structure. If

the objective in mind is to compare the regulatory networks inferred in each condition, this is a momentous drawback. In the context of systems biology, the identification of modifications in the regulatory mechanisms between different conditions is often the very purpose of the experimental design. We can think of case/control studies comparing regulatory mechanisms in diseased patients and sane controls, placebo/treatment experiments analysing the effect of a specific treatment on regulatory mechanisms compared to the placebo group. Stress experiments can lead to a large variety of gene expression profiles. More generally, even when the experimental design does not define prior sets of conditions, any known partition of some phenotypes can define *a posteriori* as many interesting sets of conditions. In all these settings, differential analyses are often led to identify the univariate variations in gene expression profiles that best distinguish those conditions. In a similar way, GGMs can be used to identify variations in conditional dependency structures across distinct conditions, hopefully unveiling changes in gene regulation mechanisms. This is the ambition of the regularizations linked to assumptions (C3) and (C4), namely the sparse group-Lasso and the cooperative-Lasso.

### 3.4.2 Illustration on Real Datasets

A thorough simulation study has been conducted by Chiquet et al. (2011), we therefore refer to this paper for numerical experiments comparing the performances of the cooperative-Lasso and group-Lasso in terms of edge detections under various settings.

In this subsection, we illustrate the benefits of the cooperative-penalty on two datasets, first of all a multiple sclerosis dataset issued from a collaboration with J.C. Corvol which led to the presentation of a poster at the European Comitee for Treatment and Research Multiple Sclerosis (ECTRIMS 2012) and secondly, a cancer dataset kindly provided by M. Jeanmougin. In both cases, the objective of the experiment is to compare the gene regulatory networks of two subpopulations.

**Multiple Sclerosis Dataset.** Gene expression profiles were taken from 26 patients with secondary-progressive multiple sclerosis (MS) included in a placebo-controlled, multiple-ascending dose, double-blind study (Kovalchin et al. 2010). Measurements were taken at baseline and once a month over the next three months. Among them, 19 patients (*active* group) were administered doses of amino acid copolymer PI-2301, which is envisaged as an alternative therapy for MS, while 7 patients (*placebo* group) received a placebo. We adopt a candidate gene approach and infer a Gaussian graphical model on 23 genes known or suspected to be genetically associated with MS.

The specificity of this dataset is that it is longitudinal. We therefore combine the VAR1 modeling and cooperative penalty. In order to correct for patient-specific effect, the data is centered and scaled patient by patient. The networks presented in Figure 3.12 were selected according to the BIC criterion.

The networks seem to share some of the paths, organized around IL2RG and IL7, but disagree on the activation of JAK1 regulations. The

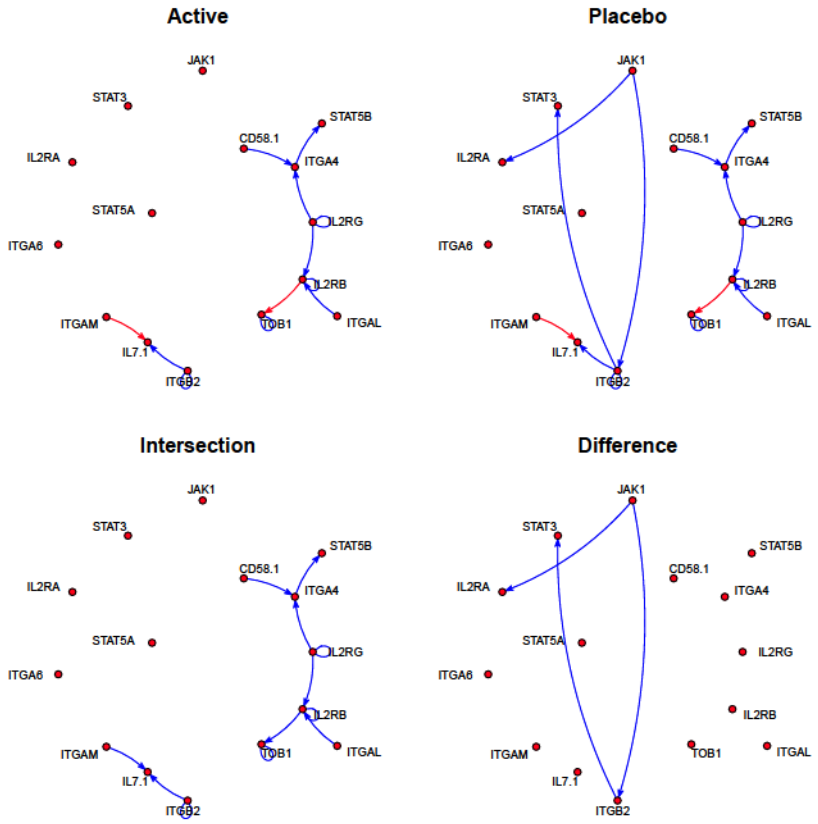



Figure 3.12 – *Gaussian graphical models inferred by the coop-Lasso on active (top-left panel) and placebo (top-right panel) groups, common (bottom-left) and condition-specific (bottom-right) edges. The amount of penalty is chosen by BIC.*

main question is whether this discrepancy is merely an estimation artefact, or if this discrepancy is statistically significant and could be interpreted as a real (potentially indirect) inhibition of JAK1 regulations by the administration of the drug.

**Cancer Relapses Dataset.** The dataset consists in 82 transcriptomes from patients suffering from breast cancer, extracted from the study conducted by Guedj et al. (2012). Patients are split into two subpopulations: 31 of them suffered from metastatic relapses (*notRFS* group), 51 did not (*relapse-free survival* (RFS) group). We restrict ourselves to the analysis of a signature of 62 genes selected by M. Jeanmougin using the approach described in Jeanmougin (2012). The networks selected by BIC criterion are presented in Figure 3.13.

Among the bulk of edges, only one (CRAP2-MMP1) happens to differ between the two sets of patients. Again, the question is whether this discrepancy is significant or not.

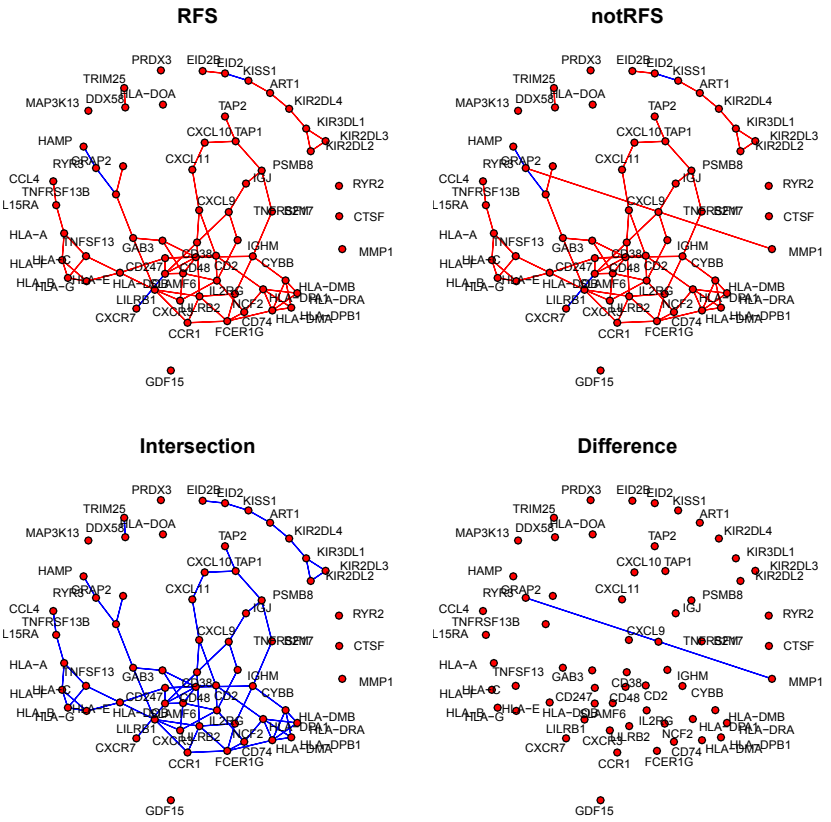



Figure 3.13 – *Gaussian graphical models inferred by the coop-Lasso on RFS (top-left panel) and notRFS (top-right panel) groups, common (bottom-left) and condition-specific (bottom-right) edges. The amount of penalty is chosen by BIC.*

# 4 Homogeneity Tests for High-Dimensional Linear Regression

This chapter presents some ongoing work in collaboration with F. Villers and N. Verzelen in an attempt at providing two-sample homogeneity tests for high-dimensional linear regression, hence the heterogeneity in the depth of analysis of the various methods explored.
We study an adaptation of the one-sample testing procedure described in Verzelen and Villers (2010) to the two-sample framework, including theoretical controls on type-I error and power.
We also include a more recent and less advanced investigation of an adaptation of higher-criticism to the two-sample testing problem, which we think particularly interesting in terms of computing time when facing high-dimensional dataset.
We provide numerical experiments illustrating the performances of those testing strategy in a rather simple design setting. We hope to gather soon some results under more complex designs.

## 4.1 Introduction

As exposed in previous chapters, the recent flood of high-dimensional data has motivated the development of a vast range of sparse estimators. If theoretical guarantees have been provided in terms of prediction, estimation and selection performances (among a lot of others Bickel et al. 2009, Wainwright 2009a, Meinshausen and Yu 2009), only a rather small proportion of the research effort focuses on quantifying the uncertainty surrounding the estimate *on a given data set with given design proportions*, be it in terms of confidence intervals or parametric hypothesis testing schemes guaranteeing a control on type I errors. Yet, quantifying the uncertainty is essential in applications where further experiments or developments rely on selected models and estimated coefficients.

This chapter is mainly motivated by the validation of differences observed between Gaussian graphical models inferred on transcriptomic data from two subpopulations, as many potentially new drug or knockout targets. Of course, graph theory comes with a vast literature about graph comparisons. Yet, we would like to stress here that our objective is not to compare two graphical structures *taken for granted*, but to test whether the divergences in estimated graphical structures could come from *estimation uncertainties*. Following literature terms, we identify the two subpopulations as two *samples*, but the reader might keep in mind that those two samples can also be referred to as conditions or tasks, as in Chapter 3, depending on the research field.

In the sequel, we keep this motivation in mind but adopt the high-dimensional linear regression model as theoretical framework. Formally, we consider the following statistical model

$$\begin{aligned} Y^{(1)} &= X^{(1)}\beta^{(1)} + \epsilon^{(1)} \\ Y^{(2)} &= X^{(2)}\beta^{(2)} + \epsilon^{(2)} , \end{aligned} \tag{4.1}$$

where the size $p$ row vectors $X^{(1)}$ and $X^{(2)}$ follow Gaussian distributions $\mathcal{N}(0_p, \Sigma^{(1)})$ and $\mathcal{N}(0_p, \Sigma^{(2)})$, whose covariance matrices remain unknown. The noise components $\epsilon^{(1)}$ and $\epsilon^{(2)}$ are independent from the design matrices and follow a centered Gaussian distribution with unknown standard deviations $\sigma^{(1)}$ and $\sigma^{(2)}$.

The objective is to test whether the models (4.1) and (4.2) are the same, that is

$$\mathcal{H}_0 : \ \beta^{(1)} = \beta^{(2)} , \quad \sigma^{(1)} = \sigma^{(2)} , \quad \text{and} \quad \Sigma^{(1)} = \Sigma^{(2)} .$$

This problem also amounts to test whether $(Y^{(1)}|X^{(1)}) \sim (Y^{(2)}|X^{(2)})$ a.s. and $\Sigma^{(1)} = \Sigma^{(2)}$.

**Remark 4.1** *The addition of $\Sigma^{(1)} = \Sigma^{(2)}$ to the null hypothesis is arguable, and clearly depends on the random design assumption. This choice comes from our motivation to derive tests for Gaussian graphical models.*

### 4.1.1 Literature in Close Frameworks

The literature on high-dimensional two-sample tests being very light, most of the state-of-the-art concerns two close subjects: high-dimensional tests for the equality of means and high-dimensional linear regression tests for the nullity of coefficients. Since tests for the equality of means corresponds to linear regression designs where $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are both equal to the identity matrix $\mathbf{I}$, the objective is to extend the former to non-orthogonal designs, exploiting ideas from the latter. Besides, since our main motivation is to deal with Gaussian graphical models, we need to find out a testing strategy adapted to random design regression.

In both scenarios as in high-dimensional estimation, the fundamental key to high-dimension is the assumption of *sparsity*. How to introduce sparsity and exploit it distinguishes the different approaches.

**Tests for the Equality of Means.** In the classical $\min(n_1, n_2) > p$ framework, it is natural to test for the equality of means via the multivariate form of the student t-statistic, the Hotelling $T^2$ statistic. Denote respectively by $\bar{\mathbf{Y}}^{(i)}$ and $\widehat{\Sigma}$ the empirical mean of the $\mathbf{Y}^{(i)}$ and common empirical covariance. The Hotteling statistic is defined by:

$$T^2 = \frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{Y}}^{(1)} - \bar{\mathbf{Y}}^{(2)})^\intercal \Sigma_n (\bar{\mathbf{Y}}^{(1)} - \bar{\mathbf{Y}}^{(2)}),$$

which follows a Fisher distribution with parameters $p$ and $n_1 + n_2 - p - 1$ under the null hypothesis. However, as underlined by Bai and Saranadasa (1996), the asymptotic power of the Hotelling test statistic suffers from the inaccurate estimation of $\Sigma$ when $p$ is of the order of $\min(n_1, n_2)$. As a result, the main challenge is to improve the inference of the common covariance matrix $\Sigma$ under difficult design sizes. Some results suggest to rely on diagonal estimates, like Bai and Saranadasa (1996), Chen and Qin (2010), Srivastava and Du (2008). Yet, the gain in power due to the faster convergence of $\widehat{\Sigma}$ is made at the price of the omition of potential correlations in the design. More recently, to refine the estimation of the full covariance matrix, Lopes et al. (2011) make use of sparsity assumptions and compute repeatedly the Hotelling test statistic on small random subsets of variables, before taking the average over all random subsets. This random projection method achieves greater power than methods based on diagonal estimates of $\Sigma$ like Bai and Saranadasa (1996), Chen and Qin (2010), Srivastava and Du (2008) as soon as variables are correlated and most of the variance can be captured by small subsets of variables.

**Tests for the Nullity of Coefficients.** Tests for the nullity of coefficients in high-dimensional linear regression form a one-sample analog of our problem. It can be considered as a limit of the two-sample test in the case where $\beta_2$ is known and equal to 0, and the sample size $n_2$ is considered infinite so that we perfectly known the distribution of the second sample. A first series of papers provide high-dimensional p-values for the nullity of coefficients in the high-dimensional linear regression framework $\mathcal{H}_{0,i} : \boldsymbol{\beta}_i^{(1)} = 0$, in the objective of testing for the significance of each coefficient individually.

Despite the problem of fitting a model with more variables than observations, which can be solved using regularized regression, the main problem in high-dimensional linear regression is to correctly estimate the variance and covariance components, just as in the test for the equality of means. Following the enthusiasm for $\ell_1$ regularized least squares, there has been attempts at providing confidence intervals for the Lasso through an estimation of the standard errors. Tibshirani (1996) adresses this issue but suggests an estimator of the standard error which inappropriately gives a null variance for all coefficients which are set to zero. Osborne et al. (2000) provide a new approximation which corrects this problem but cannot be used when the number of variables exceeds the sample size. Besides, they raise the question of whether the uncertainty surrounding Lasso coefficients can be adequately summarized by standard errors, since their distribution is likely to be distorted around zero. A somewhat answer to this issue would be provided by Bayesian approaches like the Bayesian Lasso Kyung et al. (2010), which provides posterior credible intervals for each coefficient.

In order to overcome the burden of dimension, another line of work adopts a two-step approach through half-sampling. Indeed, Wasserman and Roeder (2009) suggests to split the sample in half and apply model selection on the first half in order to test for the significance of each coefficient using the usual combination of ordinary least squares and Student t-test on a model of reasonnable size on the second half. To reduce the dependency of the results to the splitting, Meinshausen et al. (2009) advocate to use half-sampling $B$ times, and aggregate the $B$ $p$-values obtained for variable $j$ in a way which controls either the family-wise error rate or false discovery rate. They note that both methods require a $\beta_{\min}$ condition to guarantee that relevant covariates enter the model in the first step. Yet, the main issue with the procedure based upon half-sampling is that the cost of splitting the sample in half is paid twice: first, the model selection step lacks in robustness, second, the testing step is rendered strongly conservative.

The last two pieces of work manage to get rid of the $\beta_{min}$ assumption. They start from a regularized regression and build compent-wise confidence intervals or p-values for regularized estimates once corrected for bias. The approach by Zhang and Zhang (2011) provide in a way an answer to the question of confidence intervals based upon the Lasso. They define a Low-Dimensional Projection Estimator, following the efficient score function approach from semi-parametric statistics. Under classical restricted eigenvalue assumptions to guarantee the convergence of the initial Lasso estimate, as well as assumptions linked to the scaled Lasso (Antoniadis 2010, Sun and Zhang 2010; 2011) to guarantee a consistent estimation of the noise variance, they provide robust confidence intervals for each individual component of $\beta$. Bühlmann (2012b) develop a similar idea, building upon the Ridge estimator. Under mild conditions on the design, this work derives component-wise confidence intervals based upon stochastic upper-bounds for bias-corrected Ridge estimates. However, it seems from their simulated experiments that the use of stochastic upper-bounds results in a highly conservative type-I error control.

The second work also leads to p-values for tests of joint nullity

$\mathcal{H}_0^{\mathcal{S}} : \beta^\star_{\mathcal{S}^c} = 0$, for a given subset of variables $\mathcal{S} \subseteq \{1, \dots, p\}$. This testing scheme allows to test whether the true model lies within a given subspace, or whether some important variables are missing. This is the point of view adopted by Verzelen and Villers (2010). Yet in high-dimension, if it is possible to compute a sparse enough model under the null hypothesis, the full alternative model is still intractable by usual least squares. The idea of Verzelen and Villers (2010), based upon the work of Baraud et al. (2003), is to approximate the alternative $\mathcal{H}_1^{\mathcal{S}} : \exists j \in \mathcal{S}^c,\ \beta_j^\star \neq 0$ by a collection of tractable alternatives $\{\mathcal{H}_1^{\mathcal{S},m} : \exists j \in m \subset \mathcal{S}^c,\ \beta_j^\star \neq 0, m \in \mathcal{M}\}$ working on models $m$ of reasonable size. The null hypothesis is rejected if the null hypothesis $\mathcal{H}_0^{\mathcal{S}} : \beta^\star$ is rejected against at least one of the alternatives $\mathcal{H}_1^{\mathcal{S},m}$ at levels corrected for multiple testing. This approach is in a way analog to Lopes et al. (2011), since a collection of reduced models is used to approximate an untractable high-dimensional statistic.

If global testing approaches like in Verzelen and Villers (2010) is clearly less informative than approaches providing individual significance tests like Meinshausen et al. (2009), Zhang and Zhang (2011), Bühlmann (2012b), global approaches can reach better performances for fewer sample sizes. A typical and now popular example of this phenomenon is given by higher-criticism, which is known to reach optimal rates for the detection of rare and weak signals for an extremely competitive computing time. Higher-criticism was originally introduced in orthonormal designs (Donoho and Jin 2004, Hall and Jin 2008), but has been proved to reach optimal detection rates in high-dimensional linear regression as well (Arias-Castro et al. 2011, Ingster et al. 2010). In the end, higher-criticism is certainly highly competitive in terms of computing time, but requires strong assumptions on the design. The strategy adopted in Verzelen and Villers (2010) is indeed much more intensive in terms of computing, but corresponding theoretical results remain valid without any assumptions on the design.

Last but not least, we need mention a quite popular solution among biologists because of its flexibility, which is to run permutation tests. Sample indices are permuted $N$ times in order to create $N$ fictive samples $(\mathbf{X}_\pi^{(1)}, \mathbf{X}_\pi^{(2)})$ mimicking the null hypothesis that the two samples come from the same distribution. The statistic of interest $T(x^1, x^2)$ is computed on each permuted sample in order to simulate its distribution under the null. It then suffices to compare the observed statistic $T(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ to its permuted quantile $\hat{q}_\alpha(T(\mathbf{X}_\pi^{(1)}, \mathbf{X}_\pi^{(2)}))$. However efficient empirically, permutation tests present two main drawbacks beyond their computational cost. First, if the strategy seems quite natural, it is not as trivial to justify them from a theoretical point of view. Second, performances rely on the crucial identification of a relevant statistic $T(x^1, x^2)$, which is again not as trivial as it might seem.

### 4.1.2 Suggested Approach.

We recall that our main objective is to test for the homogeneity of sample-specific coefficients $\beta^{(1)} = \beta^{(2)}$ in design proportions such that the estima-

tion of high-dimensional parameters is only accessible via biased regularized estimators, whose variance terms are even harder to estimate.

To answer this question, we build upon the procedure of Verzelen and Villers (2010). The idea is to project the main statistical testing problem onto a collection of subspaces of lower dimension, and combine the results of low dimensional tests by multiple testing calibrations. In order to adapt this approach, we need to precise three steps: first, the construction of a good parametric statistic to run our tests in low dimension, second, the selection of a good collection of low dimension subspaces, third, the choice of an efficient calibration procedure. Here is a short overview of our answers to these three points.

**Choice of a Good Parametric Statistic.** Note that the two-sample test problem can be rephrased as a one sample test problem using the notation

$$\begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} & 0 \\ 0 & \mathbf{X}^{(2)} \end{bmatrix} \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \end{bmatrix} . \tag{4.2}$$

Under this formulation, it appears quite clearly that a simple Fisher statistic testing for linear constraints $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$ can be used under classical design proportions, that is to say, provided the least square estimator is defined. Therefore, a first naïve option is to combine the procedure of Verzelen and Villers (2010) with classical Fisher statistics. However, more than testing $\mathcal{H}_0$, the Fisher statistics is testing for $\beta^{(1)} = \beta^{(2)}$ under the assumption that $\sigma^{(1)} = \sigma^{(2)}$ , and $X^{(1)} \sim X^{(2)}$. In order to test $\mathcal{H}_0$ a bit more accurately and be able to reject the null if $\sigma^{(1)} \neq \sigma^{(2)}$ or $X^{(1)} \not\sim X^{(2)}$, we introduce a new likelihood-ratio-type statistic quantifying roughly-speaking how much the sample-specific estimates are far from adequately fitting the opposite sample. We prove that the procedure achieves optimal rates in the minimax sense, adapting itself to the unknown sparsity of the difference $\beta^{(1)} - \beta^{(2)}$, under minimal assumptions on the maximal sparsity to remain tractable.

**Selection of a Powerful Collection of Models.** The ideal collection models or subspaces in which to lead tests in small dimension must be exhaustive enough, so that we miss none of the most informative models. Yet, the computing time being linear in the size of the collection, we need to keep the number of models reasonably low. In order to find the best tradeoff between those two competing objectives, we investigate the use of deterministic as well as data-driven collections of models.

**Multiple-Testing Calibration.** The fine tuning of this last step is crucial to maintain a strong control on type-I Error without compromising the power of the test. We first focus on Bonferroni calibration for its simplicity of implementation as well as to derive fine theoretical controls on type-I error and power. Yet, as explained in more detail in Section 4.2, the strategy used to control the quantiles of the suggested likelihood-ratio statistic makes Bonferroni calibration even more conservative than usual. Therefore we also investigate the performances of a calibration by permutation, which admittedly takes a lot more computing time, but achieves greater power.

**Quick Look at Two-Sample Higher Criticism** Given the detection performances proved for higher-criticism in high-dimensional linear regression for such a competitive computing time, we investigate the possible adaptation of higher-criticism to the two-sample test problem. Yet, the theoretical side of two-sample higher-criticism is still to be explored.

After a short clarification of the notations, we devote Section 4.2 to the description of the adaptive likelihood-ratio procedure, along with theoretical controls of type-I error and power. Section 4.3 defines higher-criticism and explores its possible adaptation to two-samples tests. Section 4.4 provides simulated experiments comparing the performances of the suggested procedures. Section 4.6 provides additional details about the technique used in Section 4.2 to control the quantiles of the likelihood-ratio statistic.

### 4.1.3 Notation

We mention here some notation to used throughout the chapter. We consider an $n_1$-sample of the first model and an $n_2$-sample of the second model. In the sequel, the size $n_1$ (resp. $n_2$) vector of the responses $Y^{(1)}$ (resp. $Y^{(2)}$) is denoted $\mathbf{Y}^{(1)}$ (resp. $\mathbf{Y}^{(2)}$). Similarly, the design of size $n_1 \times p$ and $n_2 \times p$ are denoted $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Moreover, $\mathbf{Y}$ and $\mathbf{X}$ respectively stand for the concatenation of $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ and of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$.

$$\begin{aligned} \mathbf{Y}^{(1)} &= \mathbf{X}^{(1)}\beta^{(1)} + \boldsymbol{\epsilon}^{(1)} \\ \mathbf{Y}^{(2)} &= \mathbf{X}^{(2)}\beta^{(2)} + \boldsymbol{\epsilon}^{(2)} . \end{aligned} \tag{4.3}$$

Also, $\mathcal{L}^{(1)}$ (resp. $\mathcal{L}^{(2)}$) denotes the log-likelihood of the first (resp. second) sample normalized by $n_1$ (resp. $n_2$). Given a subset $S \subset \{1, \dots, p\}$ of size smaller than $n_1 \wedge n_2$, $(\widehat{\beta}_S^{(1)}, \widehat{\sigma}_S^{(1)})$ stands for the maximum likelihood estimator of $(\beta^{(1)}, \sigma_1)$ with the constraint that the support of $\widehat{\beta}_S^{(1)}$ is included in $S$. Similarly, we note $(\widehat{\beta}_S^{(2)}, \widehat{\sigma}_S^{(2)})$ for the maximum likelihood corresponding to the second sample, and $(\widehat{\beta}_S, \widehat{\sigma}_S)$ the maximum likelihood corresponding to the constrained model, pooling the two samples into one. Similarly, $\beta_S^{(1)}$ and $\beta_S^{(2)}$ denote the restriction of $\beta^{(1)}$ and $\beta^{(2)}$ to model $S$.

In general, respecting the notations already adopted in previous chapters, $\ell_p$ norms are denoted $\|\cdot\|_p$, except for the Euclidean norm which is sometimes referred as $\|.\|$ to alleviate notations. For any positive definite matrix $\Sigma$, $\|.\|_\Sigma$ denotes the Euclidean norm associated with the scalar product induced by $\Sigma$: for every vector $x$, $\|x\|_\Sigma = x^\intercal \Sigma x$. Besides, for every set $S$, $|S|$ denote its cardinality. For any integer $k$, $\mathbf{I}_k$ stands for the identity matrix of size $k$. For any square matrix $A$, $\varphi_{\max}(A)$ and $\varphi_{\min}(A)$ denote respectively the maximum and minimum eigenvalues of $A$. When the context makes it obvious, we may omit to mention $A$ to alleviate notations and use $\varphi_{\max}$ and $\varphi_{\min}$ instead.

To finish with, $L$ refers to a positive numerical constant that may vary from line to line.

## 4.2 Adaptive Homogeneity Tests

The overall testing scheme adopted in this section is based upon the high-dimensional parametric testing procedure described in Verzelen and Villers (2010), itself adapted from the general scheme designed by Baraud et al. (2003) in order to derive statistical tests against non-parametric alternatives. The testing procedure approximates the untractable high-dimensional test of $\mathcal{H}_0 : \beta^{(1)} = \beta^{(2)}$ against $\mathcal{H}_1 : \beta^{(1)} \neq \beta^{(2)}$ by a multiple testing construction. The approximation relies on the fundamental assumption that the true model is sparse and lies in a subspace of reasonnable dimension, compared to the sample sizes $n_1$ and $n_2$. If $S$ stands for any subset of $\{1, \dots, p\}$ that satisfies $2|S| \leq n_1 \wedge n_2$, we approximate the test of $\mathcal{H}_0$ against $\mathcal{H}_1$ by a collection of tests $\{\mathcal{H}_{0,S} \text{v.s.} \mathcal{H}_{1,S}\}_{S \in \mathcal{S}}$ in reduced dimension:

$$\begin{cases} \mathcal{H}_{0,S} : & \beta_S^{(1)} = \beta_S^{(2)}, \quad \sigma^{(1)} = \sigma^{(2)}, \quad \text{and} \quad \Sigma_S^{(1)} = \Sigma_S^{(2)}, \\ \mathcal{H}_{1,S} : & \beta_S^{(1)} \neq \beta_S^{(2)}, \quad \text{or} \quad \sigma^{(1)} \neq \sigma^{(2)}. \end{cases}$$

**Lemma 4.1** *The hypothesis $\mathcal{H}_0$ implies $\mathcal{H}_{0,S}$ for any subset $S \subset \{1, \dots p\}$.*

*Proof.* Under $\mathcal{H}_0$, the random vectors of size $p+1$ $(Y^{(1)}, X^{(1)})$ and $(Y^{(2)}, X^{(2)})$ follow the same distribution. Hence, $(Y^{(1)}|X_S^{(1)}) \sim (Y^{(2)}|X_S^{(2)})$ for any subset $S$. □

By contraposition, it suffices to reject at least one of the $\mathcal{H}_{0,S}$ hypothesis to reject the global null. Obviously, it would not be reasonable in terms of algorithm complexity to test for each null hypothesis of reduced dimension, since there would be $2^p$ of them. As a result, we must restrain ourselves to a relevant reduced collection of tests $\{\mathcal{H}_{0,S}, S \in \mathcal{S}\}$, where the collection of support $\mathcal{S}$ is potentially data-driven. On the one hand, by adding a calibration for multiple testing, we can guarantee a control on type I error. On the other hand, if the collection $\mathcal{S}$ is judiciously selected, then we can manage not to lose too much power compared to the full deterministic collection.

This framework can be described through three major steps:

1. a parametric statistic for the tests of reduced hypotheses $\mathcal{H}_{0,S}$;
2. a powerful data-driven collection of models $\widehat{\mathcal{S}}$;
3. a calibration procedure guaranteeing the control on type I error.

The next three sections discuss interesting options for these three steps.

### 4.2.1 Parametric Test Statistic

**Naïve Fisher Statistic.** For a given model $S$ of reasonable size $|S| \leq D_{max} = (n_1 \wedge n_2)/2$, testing $\mathcal{H}_{0,S}$ against the specific alternative that

$$\mathcal{H}_{1,S} : \beta_S^{(1)} \neq \beta_S^{(2)}, \quad \text{but} \quad \sigma^{(1)} = \sigma^{(2)} \quad \text{and} \quad X_S^{(1)} \sim X_S^{(2)},$$

one can naturally rely on a usual Fisher statistic testing for linear constraints $\beta_S^{(1)} = \beta_S^{(2)} = \beta_S$ on regression coefficients of model (4.2). Because

the two samples are independent, the residual sum of squares of the unconstrained model (4.2) decomposes into the sums of squares of the two sample-specific models. The Fisher statistic, with $(p, n-2p)$ degrees of freedom reads:

$$Fi_S = \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\beta}_S\|^2 - \|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2 - \|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2}{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2 + \|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2} \frac{n-2p}{p}. \quad (4.4)$$

**Likelihood-Ratio Statistic.** However, if we keep in mind our objective to derive homogeneity tests for GGMs, the assumptions that $\sigma^{(1)} = \sigma^{(2)}$ and particularly that $\Sigma^{(1)} = \Sigma^{(2)}$ can be overly restrictive, which leads us to introduce a new parametric statistic taking the form of a two-sample likelihood-ratio, measuring how far the sample-specific estimates disagree with the opposite sample. To do so, let us define the likelihood ratio at $(\beta, \sigma)$ with respect to sample $i = 1, 2$ as

$$\mathcal{D}_{n_i}^{(i)}(\beta, \sigma) := \mathcal{L}_{n_i}^{(i)}\left(\widehat{\beta}_S^{(i)}, \widehat{\sigma}_S^{(i)}\right) - \mathcal{L}_{n_i}^{(i)}(\beta, \sigma).$$

We can now consider the following statistic:

$$F_S = 2\left[\mathcal{D}_{n_1}^{(1)}(\widehat{\beta}^{(2)}, \widehat{\sigma}^{(2)}) + \mathcal{D}_{n_2}^{(2)}(\widehat{\beta}^{(1)}, \widehat{\sigma}^{(1)})\right]. \quad (4.5)$$

The statistic $F_S$ amounts to comparing the estimators $(\widehat{\beta}_S^{(1)}, \widehat{\sigma}_S^{(1)})$ and $(\widehat{\beta}_S^{(2)}, \widehat{\sigma}_S^{(2)})$ through their corresponding log-likelihoods. In order to simplify the analysis, we decompose the test statistic $F_S$ into the sum of three terms $F_{S,1} + F_{S,2} + F_{S,3}$, where

$$\begin{aligned}
F_{S,1} &= -2 + \frac{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2/n_1}{\|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2/n_2} + \frac{\|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2/n_2}{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2/n_1} \\
F_{S,2} &= \frac{\|\mathbf{X}_S^{(2)}(\widehat{\beta}_S^{(1)} - \widehat{\beta}_S^{(2)})\|^2/n_2}{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2/n_1} \\
F_{S,3} &= \frac{\|\mathbf{X}_S^{(1)}(\widehat{\beta}_S^{(1)} - \widehat{\beta}_S^{(2)})\|^2/n_1}{\|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2/n_2}.
\end{aligned}$$

While $F_{S,1}$ evaluates the discrepancies in terms of conditional variances, $F_{S,2}$ and $F_{S,3}$ compare $\beta^{(1)}$ to $\beta^{(2)}$. Proposition 4.2 characterizes the distribution of each of these terms. To simplify notations, let us denote by $g$ the non-negative function defined on $\mathbb{R}^+$ mapping $x$ to $-2 + x + 1/x$.

**Proposition 4.2** (Conditional distributions of $F_{S,1}$, $F_{S,2}$ and $F_{S,3}$ under $\mathcal{H}_0$)

1. *Let $Z$ denote a Fisher random variable with $(n_1 - |S|, n_2 - |S|)$ degrees of freedom. Then, under the null hypothesis,*

$$F_{S,1}|\mathbf{X}_S \underset{\mathcal{H}_0}{\sim} g\left[Z\,\frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)}\right].$$

2. *Let $Z_1$ and $Z_2$ be two centered and independent Gaussian vectors with covariance $\mathbf{X}_S^{(2)}\left[(\mathbf{X}_S^{(1)*}\mathbf{X}_S^{(1)})^{-1}+(\mathbf{X}_S^{(2)*}\mathbf{X}_S^{(2)})^{-1}\right]\mathbf{X}_S^{(2)*}$ and $I_{n_1-|S|}$. Then, under the null hypothesis,*

$$F_{S,2}|\mathbf{X}_S \underset{\mathcal{H}_0}{\sim} \frac{\|Z_1\|^2/n_2}{\|Z_2\|^2/n_1}.$$

*A symmetric result holds for $F_{S,3}$.*

In order to calibrate a multiple testing procedure based on these parametric statistics, we shall compute the corresponding $p$-values. Although the distributions identified in Proposition 4.2 are not familiar distributions with ready-to-use quantile tables, they all share the advantage that they do not depend on any unknown quantity, such as design variances $\Sigma^{(1)}$ and $\Sigma^{(2)}$, noise variances $\sigma^{(1)}$ and $\sigma^{(2)}$, or even true signals $\beta^{(1)}$ and $\beta^{(2)}$. In the sequel, we note $\overline{Q}_{1,|S|}(u|\mathbf{X}_S)$ (resp. $\overline{Q}_{2,|S|}(u|\mathbf{X}_S)$ and $\overline{Q}_{3,|S|}(u|\mathbf{X}_S)$) for the conditional probability that $F_{S,1}$ (resp. $F_{S,2}$ and $F_{S,3}$) is larger than $u$. Consider some $0<x<1$. By Proposition 4.2, the quantile $\overline{Q}^{-1}_{1,|S|}(x|\mathbf{X}_S)$ is easily computed analytically as a function of the quantile of a Fisher distribution. Since the conditional distribution of $F_{S,2}$ given $X_S$ only depends on $|S|$, $n_1$, $n_2$, and $\mathbf{X}_S$, one could compute $\bar{Q}_2(u|X_S)$ by Monte-Carlo simulations. However, this approach is computationally prohibitive for large collections subsets S. This is why we shall use an explicit upper bound of $\overline{Q}_{2,|S|}(u|\mathbf{X}_S)$ based on Laplace method, as given by Proposition 4.3.

**Proposition 4.3** (Upper-bound on $F_{S,2}$ and $F_{S,3}$ quantiles) *Let us note $a=(a_1,\dots,a_{|S|})$ the positive eigenvalues of*

$$\frac{n_1}{n_2(n_1-|S|)}\mathbf{X}_S^{(2)}\left[(\mathbf{X}_S^{(1)*}\mathbf{X}_S^{(1)})^{-1}+(\mathbf{X}_S^{(2)*}\mathbf{X}_S^{(2)})^{-1}\right]\mathbf{X}_S^{(2)*}.$$

*For any $u>\|a\|_1$, take*

$$\widetilde{Q}_{2,|S|}(u|\mathbf{X}_S):=\exp\left[-\frac{1}{2}\sum_{i=1}^{|S|}\log\left(1-2\lambda^* a_i\right)-\frac{n_1-|S|}{2}\log\left(1+\frac{2\lambda^* u}{n_1-|S|}\right)\right],$$

*where $\lambda^*$ is explicitely defined in Section 4.6. Then, for any $u>\|a\|_1$,*

$$\overline{Q}_{2,|S|}(u|\mathbf{X}_S)\le\widetilde{Q}_{2,|S|}(u|\mathbf{X}_S).$$

To simplify notations, we also define $\widetilde{Q}_{1,|S|}$ as equal to $\overline{Q}_{1,|S|}$. This abusive notation must not mask the essential difference between $\widetilde{Q}_{1,|S|}$ and $\widetilde{Q}_{2,|S|}$ or $\widetilde{Q}_{3,|S|}$. Indeed, $\widetilde{Q}_{1,|S|}$ is the exact quantile of $F_{S,1}$ while $\widetilde{Q}_{2,|S|}$ and $\widetilde{Q}_{3,|S|}$ only are upper-bounds on $F_{S,2}$ and $F_{S,3}$ quantiles. The consequences of this asymetry in terms of calibration of the test will be adressed in Subsection 4.2.3.

### 4.2.2 Choices of Test Collections

Many collections $\mathcal{S}$ can be thought. The ideal collection $\mathcal{S}$ must satisfy the best tradeoff between the inclusion of the maximum number of relevant models $S$ and a reasonable computing time, which is linear in the size of the collection $|\mathcal{S}|$. In the following, we distinguish deterministic and data-driven collections, which we differentiate by adding a hat on data-driven collections $\widehat{\mathcal{S}}$.

**Deterministic Collections.** Among deterministic collections of tests, the most straightforward collections consist of all size-$k$ subsets of $\{1, \dots, p\}$, which we denote $\mathcal{S}_k$. This kind of family is interesting in at least two ways. First, it neglects none of the variables: we cannot miss any signal. Second, it provides collections of tests which are independent from the data, thereby reducing the risk of overfitting. However, as we allow the model size $k$ or total number of candidate variables $p$ to grow, these deterministic families can rapidly reach unreasasonble sizes. Admittedly, $\mathcal{S}_1$ always remains feasible, but reducing the search to models of size 1 can be costly in terms of power. As a variation on size $k$ models, an interesting collection in terms of theoretical developments is the collection of all models of size smaller than $k$, denoted $\mathcal{S}_{\leq k} = \bigcup_{j=1}^{k} \mathcal{S}_j$. Note that deterministic collections can also include prior information on the model, particularly if part of the model is already known. The main point is that prior information cannot have been extracted from the same dataset.

**Data-driven Collections.** In order to investigate models of varying sizes while keeping the size of the collection moderate, we suggest to derive data-driven collections of tests $\widehat{\mathcal{S}}$. The idea is to start from a deterministic family $\mathcal{S}$ and define an algorithm an algorithm mapping $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ to some data-driven collection $\widehat{\mathcal{S}} \subset \mathcal{S}$ of restricted size. In practice, we start from $\mathcal{S}_{\leq D_{\max}}$, where $D_{\max} = \lfloor (n_1 \wedge n_2)/2 \rfloor$, and derive the collection $\widehat{\mathcal{S}}$ from the Lasso regularization path of a reparametrized joint regression model, presented in Equation (4.6).

$$\begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} & -\mathbf{X}^{(2)} \end{bmatrix} \begin{bmatrix} \theta^{(1)} \\ \theta^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \end{bmatrix}. \tag{4.6}$$

In this reparametrized model, $\theta^{(1)}$ captures the mean effect $(\boldsymbol{\beta}^{(1)} + \boldsymbol{\beta}^{(2)})/2$, while $\theta^{(2)}$ captures the discrepancy between the sample-specific effect $\boldsymbol{\beta}^{(i)}$ and the mean effect $\theta^{(1)}$, that is to say $\theta^{(2)} = (\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)})/2$. Combining this reparametrization with variable selection by the Lasso, we aim to select, on the one hand, variables presenting strong common effects through $\theta^{(1)}$, on the other hand, variables presenting strong diverging effects through $\theta^{(2)}$. We denote by $\mathcal{A}^{(1)} = \{a_1, \dots, a_{D_{\max}}\}$ the $D_{\max}$-uple of the first $D_{\max}$ selected variables by order of activation (as the penalty term of the Lasso program decreases), and by $\mathcal{A}^{(2)} = \{a_1, \dots, a_{D_{\max,2}}\}$ its restriction of the variables activated within $\theta^{(2)}$, if at most $D_{\max,2}$ variables are selected within the $\theta^{(2)}$ part.

We build two families of models from this reparametrized model: first, the increasing family $\mathcal{M}_{\theta^{(2)}}$ of variables included by the Lasso in the $\theta^{(2)}$ part, by order of activation, second the increasing family $\mathcal{M}$ of variables included by the Lars algorithm, independently from its activation in the $\theta^{(2)}$ or $\theta^{(1)}$ part.

$$\mathcal{M}_{\theta^{(2)}} = \{\bigcup_{j=1}^{k} a_j;\ k = 1, \dots, D_{\max,2}\}, \qquad \mathcal{M} = \{\bigcup_{j=1}^{k} a_j;\ k = 1, \dots, D_{max}\}.$$

The justification of the first model family is that we want to focus on variables which have disagreeing effects between the two samples. However,

the divergence between effects might only appear conditionaly on other variables with similar effects, this is why the second family is chosen to include both types of variables. In the end, we consider the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$, consisting of the reunion of both model families and $\mathcal{S}_1$,

$$\widehat{\mathcal{S}}_{\text{Lasso}} = \mathcal{M} \cup \mathcal{M}_{\theta^{(2)}} \cup \mathcal{S}_1.$$

Of course, this part of the testing strategy is highly flexible: any other relevant model selection strategy can be used.

### 4.2.3 Calibration of the Testing Procedure

Suppose that we are now given a deterministic collection $\mathcal{S}$ of subsets and an algorithm mapping $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ to some data-driven collection $\widehat{\mathcal{S}} \subset \mathcal{S}$ of restricted size. The purpose of this Section is to calibrate a multiple testing procedure based on the parametric statistics $(F_{S,1}, F_{S,2}, F_{S,3})$, $S \in \widehat{\mathcal{S}}$, so that the type-I error rate remains smaller than a chosen level $\alpha$. For the sake of simplicity, we first assume that $\emptyset \notin \mathcal{S}$.

**Bonferroni Calibration (B).** The null hypothesis $\mathcal{H}_0$ is rejected when the statistic

$$T^B_{\widehat{\mathcal{S}}} := \min_{S \in \widehat{\mathcal{S}}} \min_{i \in \{1,2,3\}} \left\{ \widetilde{Q}_{i,|S|}\left(F_{S,i} | \mathbf{X}_S\right) - \alpha_S \right\} \tag{4.7}$$

is negative. The collection of weights $\{\alpha_S,\ S \in \mathcal{S}\}$ satisfies

$$\sum_{S \in \mathcal{S}} 3\alpha_S \leq \alpha\ . \tag{4.8}$$

For the collection $\mathcal{S}_{\leq k}$, a natural choice is

$$\alpha_S = \frac{\alpha}{3k} \binom{|S|}{p}^{-1}, \tag{4.9}$$

Alternatively, one can give a Bayesian flavor to the choice of the weights $\alpha_S$, $S \in \mathcal{S}$. In fact, $T^B_{\widehat{\mathcal{S}}}$ corresponds to a Bonferroni multiple testing procedure, which allows to control the size of the corresponding test, as expressed in Proposition 4.4.

**Proposition 4.4** (Size of $T^B_{\widehat{\mathcal{S}}}$) *The statistic $T^B_{\widehat{\mathcal{S}}}$ satisfies $\mathbb{P}_{\mathcal{H}_0}[T^B_{\widehat{\mathcal{S}}} < 0] \leq \alpha$.*

*Proof.* By definition, we control the deviations of $\overline{Q}_{i,|S|}$ for each $S \in \mathcal{S}$ and each $i \in \{1,2,3\}$ under $\mathcal{H}_0$.

$$\mathbb{P}_{\mathcal{H}_0}\left[\overline{Q}_{i,|S|}\left(F_{S,i} | \mathbf{X}_S\right\}) \leq \alpha_S | \mathbf{X}_S\right] \leq \alpha_S$$

Applying a union bound and integrating with respect to $\mathbf{X}$ allows to control the type I error.

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_0}[T^B_{\widehat{\mathcal{S}}} < 0] &\leq \sum_{S \in \mathcal{S}} \sum_{i=1}^{3} \mathbb{P}\left[\widetilde{Q}_{i,|S|}\left(F_{S,i} | \mathbf{X}_S\right\}) < \alpha_S\right] \\
&\leq \sum_{S \in \mathcal{S}} \sum_{i=1}^{3} \mathbb{E}_{\mathbf{X}_S}\left[\mathbb{P}\left[\widetilde{Q}_{i,|S|}\left(F_{S,i} | \mathbf{X}_S\right\}) < \alpha_S\right]\right] \\
&\leq \sum_{S \in \mathcal{S}} 3\alpha_S \leq \alpha\ .
\end{aligned}$$

□

**Remark 4.2** (Bonferroni correction on $\mathcal{S}$ and not on $\widehat{\mathcal{S}}$) *Note that even though we restrict ourselves to the collection $\widehat{\mathcal{S}}$, the Bonferroni correction must be applied to the initial deterministic collection $\mathcal{S}$ including $\widehat{\mathcal{S}}$. Indeed, if we replace the condition (4.8) by the condition $\sum_{S\in\widehat{\mathcal{S}}} 3\alpha_S \leq \alpha$, then the size of the corresponding is not constrained anymore to be smaller than $\alpha$. This is due to the fact that we use the same data set to select $\widehat{\mathcal{S}} \subset \mathcal{S}$ and to perform the multiple testing procedure. As a simple example, consider $\mathcal{S} = \mathcal{S}[1,p]$ and*

$$\widehat{\mathcal{S}} = \left\{ \arg\min_{S\in\mathcal{S}} \bigwedge_{i=1}^{3} \widetilde{Q}_{i,|S|}\left(F_{S,i}|\mathbf{X}_S\right) \right\} .$$

*Then, computing $T_\alpha^B$ is exactly equivalent to performing a multiple testing procedure on $\mathcal{S}$.*

*The same difficulty has been tackled differently by Wasserman and Roeder (2009) and Meinshausen et al. (2009). To get rid of the dependency between model selection and hypothesis testing, both papers rely on half-sampling: model selection and hypothesis testing are led on separate halves of the dataset. However, given the small number of obversations available, half-sampling suffers from an even more reduced sample size on both fronts: model selection is rendered unstable, while testing power vanishes.*

If procedure $T_{\widehat{\mathcal{S}}}^B$ is computationally and conceptually simple, the size of the corresponding test can be much lower than $\alpha$ because of three difficulties:

1. Independently from our problem, Bonferroni corrections are known to be too conservative, especially when the number of parametric tests is large.

2. As emphasized by Remark 4.2, while the Bonferroni correction needs to be based on the whole collection $\mathcal{S}$, only the statistics $(F_{S,1}, F_{S,2}, F_{S,3})$, for $S \in \widehat{\mathcal{S}}$ are considered. Provided we could afford the computational cost of testing all models within $\mathcal{S}$, this loss cannot be compensated for if we use the Bonferroni correction.

3. As underlined in Subsection 4.2.1, for computational reasons, we do not consider in (4.7) the conditional $p$-value $\overline{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S)$ and $\overline{Q}_{3,|S|}(F_{S,3}|\mathbf{X}_S)$ but only upper bounds $\widetilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S)$ and $\widetilde{Q}_{3,|S|}(F_{S,3}|\mathbf{X}_S)$ of them. We therefore overestimate the type I error due to $F_{S,2}$ and $F_{S,3}$.

We address the three aforementionned issues applying a permutation approach.

**Calibration by permutation (P).** Given a permutation $\pi$ of the set $\{1,\dots,n_1+n_2\}$, one gets $\mathbf{Y}^\pi$ and $\mathbf{X}^\pi$ by permuting the components of $\mathbf{Y}$ and the rows of $\mathbf{X}$. This allows to us to get a new sample $(\mathbf{Y}^{\pi,(1)}, \mathbf{Y}^{\pi,(2)}, \mathbf{X}^{\pi,(1)}, \mathbf{X}^{\pi,(2)})$. Using this new sample, one computes a new collection $\widehat{\mathcal{S}}^\pi$ and parametric statistics $F_{S,1}^\pi$, $F_{S,2}^\pi$, $F_{S,3}^\pi$, respectively. We note $\mathcal{P}$ the uniform distribution over the permutations of size $n_1+n_2$.

For $i \in \{1,2,3\}$, define $\widehat{C}_{i,P}$ as the $1-\alpha/3$-quantiles with respect to $\mathcal{P}$ of

$$\min_{S\in\widehat{\mathcal{S}}^{\pi}} \left\{ \widetilde{Q}_{i,|S|}\left(F^{\pi}_{S,i}|\mathbf{X}^{\pi}_{S}\right)\binom{p}{|S|} \right\}.$$

In practice, we estimate the quantiles $\widehat{C}_{i,P}$ by sampling a large number $N$ of permutations. Given $(\widehat{C}_{1,P},\widehat{C}_{2,P},\widehat{C}_{3,P})$, we build the statistic $T^{P}_{\alpha}$. The hypothesis $\mathcal{H}_0$ is rejected when the statistic

$$T^{P}_{\widehat{\mathcal{S}}} := \min_{S\in\widehat{\mathcal{S}}} \min_{i\in=\{1,2,3\}} \left\{ \widetilde{Q}_{i,|S|}\left(F_{S,i}|\mathbf{X}_S\right) - \widehat{C}_{i,P}\binom{p}{|S|}^{-1} \right\} \tag{4.10}$$

is negative. Proposition 4.5 proves that the procedure by permutation allows to control the type-I error rate at level $\alpha$.

**Proposition 4.5** (Size of $T^{P}_{\widehat{\mathcal{S}}}$) *The statistic $T^{P}_{\widehat{\mathcal{S}}}$ satisfies*

$$\alpha/3 \leq \mathbb{P}_{\mathcal{H}_0}\left[T^{P}_{\widehat{\mathcal{S}}} < 0\right] \leq \alpha\,.$$

*Proof.* Consider $i \in \{1,2,3\}$. Under $\mathcal{H}_0$, the distribution of

$$\min_{S\in\widehat{\mathcal{S}}_{\pi}} \left\{ \widetilde{Q}_{i,|S|}\left(F_{S,i}(\pi)|\mathbf{X}^{\pi}_{S}\right)\binom{p}{|S|} \right\}\,.$$

is invariant with respect to the permutation $\pi$. Hence, we derive

$$\mathbb{P}_{\mathcal{H}_0}\left[\min_{S\in\widehat{\mathcal{S}}} \overline{Q}_{i,|S|}\left(F_{S,i}|\mathbf{X}_S\right)\binom{p}{|S|} \leq \widehat{C}_{i,P}\middle|\, \mathbf{X}_S\right] = \alpha/3\,.$$

Applying a union bound and integrating with respect to $\mathbf{X}$ allows us to conclude. □

**Remark 4.3** *Through the three constants $\widehat{C}_{1,P}$, $\widehat{C}_{2,P}$ and $\widehat{C}_{3,P}$, this permutation approach corrects simultaneously for the three losses mentioned earlier due to the Bonferroni correction, the restriction to a data-driven class $\widehat{\mathcal{S}}$ and the upper bounds of $\overline{Q}_{S,2}$ and $\overline{Q}_{S,3}$.*

*Yet, the level of $T^{P}_{\widehat{\mathcal{S}}}$ is not exactly $\alpha$ because we treat separately the statistics $F_{S,1}$, $F_{S,2}$ and $F_{S,3}$ and apply a Bonferroni correction at this second level. It would be possible to calibrate all the statistics simultaneously in order to constrain the size of the corresponding test to be exactly $\alpha$. However, this last approach would favor the statistic $F_{S,1}$ too much, because we would put on the same level the true quantile $\overline{Q}_{S,1}$ and the upper bounds $\tilde{Q}_{S,2}$ and $\tilde{Q}_{S,3}$.*

### 4.2.4 Power of the Procedure

In this section, we consider some number $\delta \in (0,1)$. The objective is to prove that $T^{B}_{\widehat{\mathcal{S}}}$ reaches powers exceeding $1-\delta$ on a large set of values $(\beta^{(1)},\sigma^{(1)},\beta^{(2)},\sigma^{(2)})$ in the alternative. Because of the difficulties introduced by permutations, we are still working on the proofs concerning $T^{P}_{\mathcal{S}}$ and ultimately $T^{P}_{\widehat{\mathcal{S}}}$. For now, we rely on simulated experiments to illustrate that $T^{P}_{\mathcal{S}}$ and $T^{P}_{\widehat{\mathcal{S}}}$ actually achieve great power values.

As the analysis is more straightforward, we start by considering the power of $T^{B}_{\mathcal{S}}$ with a deterministic collection $\mathcal{S}$.

**Power of $T_S^B$ for a Deterministic Collection** Intuitively, $T_S^B$ should reject $\mathcal{H}_0$ with large probability when $(\beta^{(1)}, \sigma_1)$ is far from $(\beta^{(2)}, \sigma_2)$ in some sense. A classical way of measuring the divergence between two distributions is the Kullback-Leibler discrepancy. In the sequel, we note

$$\mathcal{K}\left[\mathbb{P}_{Y^{(1)}|X}; \mathbb{P}_{Y^{(2)}|X}\right] . \tag{4.11}$$

the Kullback discrepancy between the conditional distribution of $Y^{(1)}$ given $X^{(1)} = X$ and conditional distribution of $Y^{(2)}$ given $X^{(2)} = X$. Then, $\mathcal{K}_1$ denotes the expectation of this Kullback divergence (4.11) with respect to $X \sim X^{(1)}$. Exchanging the roles of $X^{(1)}$ and $X^{(2)}$, we also define $\mathcal{K}_2$:

$$\mathcal{K}_1 := \mathbb{E}_{X^{(1)}}\left\{\mathcal{K}\left[\mathbb{P}_{Y^{(1)}|X}; \mathbb{P}_{Y^{(2)}|X}\right]\right\} , \quad \mathcal{K}_2 := \mathbb{E}_{X^{(2)}}\left\{\mathcal{K}\left[\mathbb{P}_{Y^{(2)}|X}; \mathbb{P}_{Y^{(1)}|X}\right]\right\} .$$

The sum $\mathcal{K}_1 + \mathcal{K}_2$ forms a semidistance with respect to $(\beta^{(1)}, \sigma_1)$ and $(\beta^{(2)}, \sigma_2)$ as proved by the following decomposition

$$2\left(\mathcal{K}_1 + \mathcal{K}_2\right) = \left(\frac{\sigma^{(1)}}{\sigma^{(2)}}\right)^2 + \left(\frac{\sigma^{(2)}}{\sigma^{(1)}}\right)^2 - 2 + \frac{\|\beta^{(2)} - \beta^{(1)}\|^2_{\Sigma^{(2)}}}{(\sigma^{(1)})^2} + \frac{\|\beta^{(2)} - \beta^{(1)}\|^2_{\Sigma^{(1)}}}{(\sigma^{(2)})^2}.$$

We therefore adopt this semidistance as a measure of proximity between $(\beta^{(1)}, \sigma_1)$ and $(\beta^{(2)}, \sigma_2)$.

In the case where $X^{(1)}$ and $X^{(2)}$ do not follow the same distribution, in other words $\Sigma^{(1)} \neq \Sigma^{(2)}$, we also need to quantify the distance between their distributions, equivalently measured by the distance between the covariance matrices. Given a model $S \in \mathcal{S}$, we use the following measure of proximity between $\Sigma_S^{(1)}$ and $\Sigma_S^{(2)}$:

$$\varphi_S := \varphi_{\max}\left\{\sqrt{\Sigma_S^{(2)}}(\Sigma_S^{(1)})^{-1}\sqrt{\Sigma_S^{(2)}} + \sqrt{\Sigma_S^{(1)}}(\Sigma_S^{(2)})^{-1}\sqrt{\Sigma_S^{(1)}}\right\} .$$

First, we control the power of $T_S^B$ for a collection $\mathcal{S} = \mathcal{S}_{\leq k}$ with $k \leq (n_1 \wedge n_2)/2$ and the weights (4.9). We write $S^{\cup(1,2)}$ for the union of the supports of $\beta^{(1)}$ and $\beta^{(2)}$. Furthermore, we define $S^{\Delta(1,2)} := \{i,\ \beta_i^{(1)} \neq \beta_i^{(2)}\}$ as the set of indices such that the components of $\beta^{(1)}$ and $\beta^{(2)}$ are different.

The following proposition, which gives two different conditions under which testing procedure $T_S^B$ achieves greater power than $1 - \delta$, is a consequence of a more general result stated as Theorem 4.7. Part (a) of Proposition 4.6 provides a general condition, valid when $X^{(1)}$ and $X^{(2)}$ do not necessarily follow the same distribution. Part (b) provides a weaker condition, but which remains valid in the special case where $X^{(1)}$ and $X^{(2)}$ follow the same distribution only. Remark 4.4 gives insights into the optimality of those conditions.

**Proposition 4.6** (Power of $T_S^B$ for $\mathcal{S} = \mathcal{S}_{\leq k}$) *There exists positive numbers $L_1$ and $L_2(\alpha, \delta)$ such that the following holds. Assume that $\log(1/(\alpha\delta)) \leq L_1(n_1 \wedge n_2)$ and define $k^*$ as the largest integer satisfying*

$$k^* \log(p) \leq L_1\left(n_1 \wedge n_2\right) . \tag{4.12}$$

(a) *The hypothesis $\mathcal{H}_0$ is rejected by $T_{\mathcal{S}}^B$ with probability larger than $1-\delta$ for any $(\beta^{(1)},\beta^{(2)})$ satisfying $|S^{\cup(1,2)}| \leq k \wedge k^*$ and*

$$\mathcal{K}_1 + \mathcal{K}_2 \geq L_2(\alpha,\delta)\varphi_{S^{\cup(1,2)}} \frac{|S^{\cup(1,2)}| \vee 1}{n_1 \wedge n_2} \log(p) \ , \tag{4.13}$$

(b) *If $\Sigma^{(1)} = \Sigma^{(2)} = \Sigma$, the hypothesis $\mathcal{H}_0$ is rejected by $T_{\mathcal{S}}^B$ with probability larger than $1-\delta$ for any $(\beta^{(1)},\beta^{(2)})$ satisfying $|S^{\Delta(1,2)}| \leq k \wedge k_*$ and*

$$\frac{\|\beta^{(1)}-\beta^{(2)}\|_\Sigma^2}{\wedge_{i=1}^2 Var\left[Y^{(i)}|X^{(i)}_{S^{\Delta(1,2)}}\right]} \geq L_2(\alpha,\delta)\frac{|S^{\Delta(1,2)}| \vee 1}{n_1 \wedge n_2} \log(p) \ , \tag{4.14}$$

**Remark 4.4**

- *Assume that the true vectors are sparse, that is $|S^{\cup(1,2)}| \ll p$ and that $\varphi_{S(1,2)}$ is bounded. Then, condition (4.13) tells us that $T_{\mathcal{S}}^B$ is powerful as long as*

$$\mathcal{K}_1 + \mathcal{K}_2 \gtrsim \frac{|S^{\cup(1,2)}|}{n_1 \wedge n_2} \log(p) \ .$$

  *The rate depends on the sparsity index of the union of the supports.*

- *If $\Sigma^{(1)} = \Sigma^{(2)}$, then $T_{\mathcal{S}}^B$ is powerful as long as*

$$\frac{\|\beta^{(1)}-\beta^{(2)}\|_{\Sigma^{(1)}}^2}{\wedge_{i=1}^2 Var\left[Y^{(i)}|X^{(i)}_{S^{\Delta(1,2)}}\right]} \gtrsim \frac{|S^{\Delta(1,2)}|}{n_1 \wedge n_2} \log(p) \ . \tag{4.15}$$

  *Here, the rate only depends on $|S^{\Delta(1,2)}|$, which corresponds to the sparsity of the difference $\beta^{(1)}-\beta^{(2)}$. The cardinality $|S^{\Delta(1,2)}|$ can be much smaller than $|S^{\cup(1,2)}|$, when $\beta^{(1)}$ and $\beta^{(2)}$ share many common coefficients. In the specific case where $\beta^{(2)} = 0$, $\sigma_1 = \sigma_2$ and $n_2 = \infty$ (one-sample testing problem), (4.15) has been proved (Verzelen and Villers 2010) to be optimal in the minimax sense. In this sense $T_{\mathcal{S}}^B$ is adaptive to the unknown sparsity of the difference $\beta^{(1)}-\beta^{(2)}$.*

- *Condition (4.12) roughly tells us that the maximal sparsity $k^*$ should satisfy a condition of the type $k\log(p) \lesssim (n_1 \wedge n_2)$. This condition has been shown Verzelen (2012) to be minimal to obtain rates of testing of the form (4.13) in the specific case where $\beta^{(2)} = 0$, $\sigma_1 = \sigma_2$ and $n_2 = \infty$.*

Let us turn to a more general control of the power of $T_{\mathcal{S}}^B$ for arbitrary collections $\mathcal{S}$. To do so, we need to consider the Kullback discrepancy between the conditional distribution of $Y^{(1)}$ given $X_S^{(1)} = x_S$ and the conditional distribution of $Y^{(2)}$ given $X_S^{(2)} = x_S$, which we denote $\mathcal{K}\left[\mathbb{P}_{Y^{(1)}|X_S}; \mathbb{P}_{Y^{(2)}|X_S}\right]$. For short, we respectively note $\mathcal{K}_1(S)$ and $\mathcal{K}_2(S)$

$$\begin{aligned}
\mathcal{K}_1(S) &:= \mathbb{E}_{X_S^{(1)}}\left\{\mathcal{K}\left[\mathbb{P}_{Y^{(1)}|X_S}; \mathbb{P}_{Y^{(2)}|X_S}\right]\right\} \ , \\
\mathcal{K}_2(S) &:= \mathbb{E}_{X_S^{(2)}}\left\{\mathcal{K}\left[\mathbb{P}_{Y^{(2)}|X_S}; \mathbb{P}_{Y^{(1)}|X_S}\right]\right\} \ .
\end{aligned}$$

Intuitively, $\mathcal{K}_1(S) + \mathcal{K}_2(S)$ corresponds to some distance between the regression of $Y^{(1)}$ given $X_S^{(1)}$ and of $Y^{(2)}$ given $X_S^{(2)}$.

**Theorem 4.7** (Power of $T_{\mathcal{S}}^{B}$ for any deterministic $\mathcal{S}$) *There exist positive constants $L_1$ and $L_2(\alpha,\delta)$ such that the following holds. Consider the subcollection $\mathcal{S}' \subset \mathcal{S}$ consisting of subsets $S$ that satisfy*

$$\log[16/(\delta\alpha_S)] \;\leq\; L_1(n_1 \wedge n_2)\,. \tag{4.16}$$

*The hypothesis $\mathcal{H}_0$ is rejected by $T_{\mathcal{S}}^{B}$ with probability larger than $1-\delta$ for any $(\beta^{(1)},\beta^{(2)},\sigma_1,\sigma_2,\Sigma^{(1)},\Sigma^{(2)})$ belonging to the set*

$$\mathcal{F}_{\mathcal{S}}(\delta) := \left\{(\beta^{(1)},\beta^{(2)},\sigma_1,\sigma_2,\Sigma^{(1)},\Sigma^{(2)}),\ \exists S \in \mathcal{S}' :\ \mathcal{K}_1(S)+\mathcal{K}_2(S) \geq \Delta(S)\right\},$$

*where*

$$\Delta(S) := L_3(\alpha,\delta)\varphi_S\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left[|S|+\log(1/\alpha_S)\right]\,. \tag{4.17}$$

**Remark 4.5** *The test $T_{\mathcal{S}}^{B}$ is powerful as long as for some $S^* \in \mathcal{S}'$, $\mathcal{K}_1(S^*)+\mathcal{K}_2(S^*)$ is larger than $\Delta(S^*)$. The term $\Delta(S)$ plays the role of a variance term and increases with the cardinality $|S|$. Furthermore, the semi-distance $\mathcal{K}_1(S)+\mathcal{K}_2(S)$ has also a tendency to increase with $|S|$. Thus, $T_{\mathcal{S}}^{B}$ rejects $\mathcal{H}_0$ with large probability if the tradeoff between $-[\mathcal{K}_1(S)+\mathcal{K}_2(S)]$ and $\Delta(S)$ is negative. Our multiple testing approach allows us to reject without knowning in advance for which model $S^*$ the tradeoff is achieved. Nevertheless, we have to pay a price for this feature of adaptation: $\Delta(S)$ in Equation (4.17) becomes logarithmically larger with the size of $\mathcal{S}$ through the term $\log(1/\alpha_S)$. This phenomenon also occurs in adaptive testing in the Gaussian linear model Baraud et al. (2003).*

**Power of $T^{B}_{\widehat{\mathcal{S}}_{\text{Lasso}}}$ with the Lasso Collection $\widehat{\mathcal{S}}_{\text{Lasso}}$.** The test $T_{\mathcal{S}}^{B}$ with the collection $\mathcal{S} = \mathcal{S}_{\leq(n_1\wedge n_2)/2}$ is computationally expensive since its size is non polynomial with respect to $p$. The collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ has been introduced as a way to fix this issue, as its size is linear in $n_1 \wedge n_2$. Theorem 4.8 states that the power of $T^{B}_{\widehat{\mathcal{S}}_{\text{Lasso}}}$ is also optimal under further assumptions on the covariance matrices $\Sigma^{(1)}$ and $\Sigma^{(2)}$. In the statement below, $\psi_{\Sigma^{(1)},\Sigma^{(2)}}$ refers to a positive quantity that only depends on the largest and the smallest eigenvalues of $\Sigma^{(1)}$ and $\Sigma^{(2)}$. The expression of $\psi_{\Sigma^{(1)},\Sigma^{(2)}}$ is made explicit in the proof.

**Theorem 4.8** *There exist positive constants $L_1$, $L_2$ and $L_3(\alpha,\delta)$ such that the following holds. Assume that*

$$\log[24/(\alpha\delta)] \leq L_1(n_1 \wedge n_2).$$

*The hypothesis $\mathcal{H}_0$ is rejected by $T^{B}_{\widehat{\mathcal{S}}_{\text{Lasso}}}$ with probability larger than $1-\delta$ for any $(\beta^{(1)},\beta^{(2)})$ satisfying*

$$|S^{\cup(1,2)}| \leq L_2\psi_{\Sigma^{(1)},\Sigma^{(2)}}\frac{n_1\wedge n_2}{\log(p)}\,, \tag{4.18}$$

*and*

$$\mathcal{K}_1+\mathcal{K}_2 \geq L_3(\alpha,\delta)\psi_{\Sigma^{(1)},\Sigma^{(2)}}\frac{|S^{\cup(1,2)}|\vee 1}{n_1\wedge n_2}\log(p)\,. \tag{4.19}$$

**Remark 4.6**

- *The rates of testing* (4.19) *and sparsity condition* (4.18) *are analogous to what has been obtained in Proposition 4.6 for a deterministic collection.*
- *Dependencies of* (4.18) *and* (4.19) *on* $\Sigma^{(1)}$ *and* $\Sigma^{(2)}$ *are unavoidable because the collection* $\widehat{\mathcal{S}}_{\text{Lasso}}$ *is based on the Lasso estimators which require design assumptions to work well (Candes and Plan 2007). Nevertheless, one can improve* (4.18) *and* (4.19) *by using restricted eigenvalues instead of largest eigenvalues (See Section A.2.4 for a sharper statement).*

## 4.3 Higher-Criticism Detection of Heterogeneity

Considering the growing cases of high-dimensional samples, it is worth considering computationaly efficient testing methods like Higher-Criticism (HC). Besides, Higher-Criticism is proved optimal in terms of signal detection in one-sample linear regression by two independent and simultaneous works, Ingster et al. (2010) and Arias-Castro et al. (2011), when the signal is so sparse and weak in intensity, that usual ANOVA methods or multiple testing fail to detect it. Yet, it is often observed in genetics, be it with Single Nucleotide Polymorphism (SNP) or transcriptomic data, that among the thousand, if not millions, of candidate genes, only a few of them share tiny effects on the outcome. Higher-criticism is based on the idea that facing such stringent design proportions and rare and weak signals, under some assumptions on the design, one might not be able to identify the exact position of nonnull signal components, but might still be able to detect the presence of a nonnull signal through the detection of distorsions in the distribution of $p$-values. The objective is therefore to test for the global null hypothesis that $\mathcal{H}_0 : \beta^\star = 0$ in the one-sample linear regression scenario, without necessarily being able to identify which are the exact components being responsible for the rejection of the null.

In the sequel, we recall the principle of higher-criticism in one-sample high-dimensional linear regression before suggesting an adaptation to our two-sample testing problem.

### 4.3.1 One-Sample High-Criticism under the Rare and Weak Model

Before suggesting an adaptation of higher-criticism to the detection of differences between samples, let us recall the principle of higher-criticism in the context of one-sample linear regression. We are given a size-$n$ response vector $\mathbf{Y}$ and an $n \times p$ design matrix $\mathbf{X}$, linked through the following linear regression model : there exists a signal $\beta^\star$ and a Gaussian noise vector with unknown covariance $\sigma^2$ such that

$$Y = X\beta^\star + \varepsilon.$$

**The Rare and Weak Model.** The rare and weak model features two parameters $\eta$ and $r$, which respectively determine the sparsity and strength of the signal. The number of non zero components of $\beta^\star$ is modeled by $s = p^{1-\eta}$, $\eta \in ]0, 1[$, while all non zero components share the same value

$\mu = \sqrt{2r \log p}$, $r \in ]0,1[$. This parametrization maps two quantities depending on $p$ to the square $]0,1[\times]0,1[$.

The reasoning behind the parametrization of non zero components appears clearly if we make the simplifying assumption that the design reduces to the identity matrix $\mathbf{X} = \mathbf{I}$. In that context, keeping $\mu$ smaller than $\sqrt{2 \log p}$ ensures that in expectation, $\mu$ remains smaller than the largest $y_i$ under $\mathcal{H}_0 : \beta^\star = 0$. On the contrary, if one would allow $\mu$ to exceed $\sqrt{2 \log p}$, there would be such a distorsion of extreme values under $\mathcal{H}_1$ that the testing problem would get trivial asymptotically.

To provide some insight into the sparsity parametrization, it is useful to recall some optimality results in terms of detection boundary derived repeatedly, first on the detection of Gaussian mixtures (Donoho and Jin (2004), Cai et al. (2007), Donoho and Jin (2009), Haupt et al. (2008; 2010)) and more recently extended to linear regression under different alternatives and assumptions on the design matrix $\mathbf{X}$ (Ingster et al. 2010, Arias-Castro et al. 2011). The question of the detection boundary consists in identifying the smallest signal intensity (measured in some specific sense) such that it remains possible to detect it. It is an asymptotic version of separation distances studied in the power analysis of Section 4.2.4. A testing strategy is optimal if it is successful in detecting signals lying at the boundary. For instance, testing the joint nullity of coefficients with ANOVA is only optimal under mild levels of sparsity, $\eta \in ]0, 1/2[$, that is to say, $s \in ]\sqrt{p}, p[$. On the other side of the spectrum, multiple testing combined with a Bonferroni correction is only optimal under very strong sparsity, $\eta \in ]3/4, 1[$, or $s \in ]1, p^{1/4}[$. Only higher-criticism reaches optimality in between, for strong, but also very strong levels of sparsity, with $\eta \in ]1/2, 1]$, $s \in ]1, \sqrt{p}[$. Under the additional assumption that $p^{1-\eta} \log p = o(\sqrt{n})$, the optimal detection rate can be expressed precisely for strong levels of sparsity, coinciding in both Gaussian mixture and linear regression frameworks:

$$\rho(\eta) = \begin{cases} \eta - \frac{1}{2} & \text{if } \frac{1}{2} < \eta \leq \frac{3}{4}, \\ (1 - \sqrt{1-\eta})^2 & \text{if } \frac{3}{4} < \eta \leq 1 \end{cases}.$$

Figure 4.1 represents this phenomenon as a phase diagram, adding for comparison the estimation boundary as given in Donoho and Jin (2004).

**HC statistic.** Following Ingster et al. (2010), the higher-criticism (HC) statistic for linear regression is based upon the univariate p-values

$$q_j = \mathbb{P}\left(\mathcal{N}(0,1) \geq \frac{|\langle \mathbf{y}, \mathbf{x}_j \rangle|}{\|\mathbf{y}\|}\right), \quad j = 1, \ldots, p$$

The HC statistic is defined by as the supremum of the scaled and centered empirical process of the p-values

$$HC^\star = \sup_{0 \leq \alpha \leq \alpha_{max}} \sqrt{p} \frac{\frac{1}{p}\sum_{j=1}^{p} \mathbf{1}\left\{q_j \leq \alpha\right\} - \alpha}{\sqrt{\alpha(1-\alpha)}}.$$

For a given $\alpha$, this statistic can be understood as a second-level significance test, answering the question: are there many more significant univariate

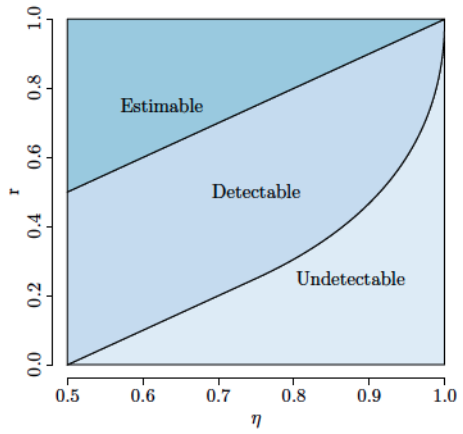



Figure 4.1 – *Phase diagram representing the detection and estimation boundaries (Donoho and Jin 2004).*

hypotheses at level $\alpha$ than merely by chance under the global null hypothesis? $HC^\star$ takes the supremum of this quantity over a range of levels $\alpha \in [0, \alpha_{max}]$.

Denoting by $q_{(j)}$ the ordered p-values, the HC statistic can be equivalently found under the following form, which is directly computable from the ordered sequence of p-values:

$$HC^\star = \sup_{j=1,\dots,p | q_{(j)} \leq \alpha_{max}} \sqrt{p} \frac{j/p - q_{(j)}}{\sqrt{q_{(j)}(1-q_{(j)})}}.$$

**Calibration.** Asymptotic theory of empirical processes gives that under the global null hypothesis

$$\frac{HC^\star}{\sqrt{2 \log\log p}} \xrightarrow{P} 1, \quad \text{when } p \to \infty.$$

This convergence in probability leads most papers (Donoho and Jin 2004, Ingster et al. 2010, Hall and Jin 2010, Arias-Castro et al. 2011) to advocate the following decision rule: reject $\mathcal{H}_0$ as soon as $HC^\star$ exceeds $(1 + a_p)\sqrt{2 \log\log p}$, with $a_p$ tending to 0. Yet this rule does not allow us to calibrate the test for a given $p$.

In the following subsection, we suggest an adaptation of HC to the detection of heterogeneity in the two-sample linear regression framework and derive a calibration based on Monte-Carlo simulations.

### 4.3.2 Two-sample Higher-criticism

**Two-sample HC statistic.** Since the principle of HC is to replace a high-dimensional multivariate statistic with the second-level analysis of the distribution of p-values testing for univariate hypotheses, any univariate statistic testing for $\beta_j^{(1)} = \beta_j^{(2)}$ has a potential interest to adapt HC to the two-sample test. In the sequel, we define the two-sample HC statistic upon the Fisher statistics testing for the equality of coefficients in each model of

size 1. Denoting by $RSS_j$ the residual sum of squares of the pooled regression of the concatenation of $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ on $\mathbf{X}_j$, the concatenation of $\mathbf{X}_j^{(1)}$ and $\mathbf{X}_j^{(2)}$, and respectively by $RSS_j^{(1)}$ and $RSS_j^{(2)}$ the residual sums of squares in sample-specific regressions, the new statistic is given by

$$z_j^F := \frac{RSS_j - RSS_j^{(1)} - RSS_j^{(2)}}{RSS_j^{(1)} + RSS_j^{(2)}} \frac{n-2}{1}.$$

Since $z_j^F$ follows a Fisher distribution with parameters $(1, n-2)$ under the null hypothesis, we can consider the p-value $q_j^F$ associated with the Fisher distribution and define the corresponding HC statistic:

$$HC^F := \sup_{j=1,\dots,p | q_{(j)}^F \leq \alpha_{max}} \sqrt{p} \frac{j/p - q_{(j)}^F}{\sqrt{q_{(j)}^F (1 - q_{(j)}^F)}}.$$

**Correcting for Common Effects.** The main disadvantage of the HC approach in the detection of heterogeneous components in the two-sample framework is that heterogeneous effets might appear only conditionaly to some common effect with respect to other features. Also, if large common coefficients are not corrected for, then their effect passes through the noise vectors $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$. As a result, larger estimated conditional variances $\widehat{\sigma}_1$ and $\widehat{\sigma}_2$ will make it more difficult to detect small heterogeneous effect as significant. Attempts at a correction for common effects will be presented in Section 4.4.

**Calibration.** All $HC$ statistics only rely on a sequence of $p$ p-values. If one can assume that those are independent from each other, then the sequence of p-values is a realization of $p$ i.i.d. uniform distributions. Therefore, for a given number of variables $p$, one can estimate quite efficiently the right-hand-side quantile of level $\alpha$ of any $HC$ statistic under $\mathcal{H}_0$. We suggest to use this quantile to calibrate the tests based on HC statistics. In practice, numerical experiments as those presented in Section 4.4 show that this calibration works well.

## 4.4 Numerical Experiments

### 4.4.1 Synthetic Linear Regression Data

**Simulation Framework.** In order to calibrate the difficulty of the testing task, we simulate our data according to the parametrization of the Rare and Weak framework presented in Section 4.3. We choose a large but still reasonable number of variables $p = 200$, but restrict ourselves to cases where the number of observations remain smaller than $p$. With equal sample sizes, we let $n_1 = n_2 = n$ take the values $n = 25,\ 50,\ 100$, and for each simulated sample, we generate two sub-samples:

$$\begin{aligned} \mathbf{Y}^{(1)} &= \mathbf{X}^{(1)}\beta^{(1)} + \varepsilon^{(1)}, \\ \mathbf{Y}^{(2)} &= \mathbf{X}^{(2)}\beta^{(2)} + \varepsilon^{(2)}. \end{aligned}$$

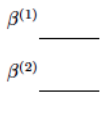
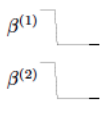
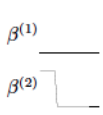
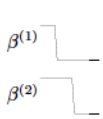
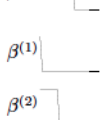
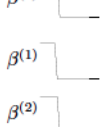
| Setting | $\eta$ | ♯ common | $\eta_2$ | ♯ $\beta^{(2)}$ specific | Signals |
|---|---|---|---|---|---|
| $\mathcal{H}_{00}$ | - | 0 | - | 0 |  |
| $\mathcal{H}_0$ | 5/8 | 7 | - | 0 |  |
| 1 | - | 0 | 5/8 | 7 |  |
| 2 | 5/8 | 7 | 5/8 | 7 |  |
| 3 | 7/8 | 1 | 5/8 | 7 |  |
| 4 | 5/8 | 7 | 7/8 | 1 |  |

Table 4.1 – *Summary of the six different simulation scenarios under study.*

In this chapter, we present preliminary results where $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are always generated under the simple scenario where observation follows a standard multivariate Gaussian distribution, $\mathbf{X}_i^{(j)} \sim \mathcal{N}(0, \mathbf{I}_p)$, and noise components $\varepsilon_i^{(1)}$ and $\varepsilon_i^{(2)}$ admit the same variances $\sigma^{(1)} = \sigma^{(2)} = 1$. We expect to gather soon some new simulation results under wider assumptions on the design matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$.

We study six different scenarios summarized in Table 4.1. The first two check for type I error control. The last four allow us to compare the performances of the various statistics under different sparsity levels and proportions of shared coefficients. These alternative scenarios are parametrized by the number of non zero common coefficients $p^{1-\eta}$, the number of non zero coefficients $p^{1-\eta_2}$ activated in $\beta^{(2)}$ only, and the magnitude $\mu = \sqrt{2r \log p}$ of all non zero coefficients. We choose the support of $\beta^{(1)}$ to be included in $\beta^{(2)}$, so that either coefficients are common to both coefficients, or they are activated in $\beta^{(2)}$ only. The Parameters $\eta$ and $\eta_2$ are chosen to generate strong and very strong levels of sparsity. The last column of Table 4.1 illustrates the signal sparsity patterns of $\beta^{(1)}$ and $\beta^{(2)}$ associated with each scenario. The two patterns of scenario 4 are so close that the illustration might be misleading: the two patterns not equal but actually differ by only one covariate. In all scenarios, the magnitude ranges from $r = 0$ to $r = 0.5$.

We repeat the experiment 1000 times, except for the case $n = 100$, for which we only gathered 500 simulations.

We start by considering the three statistics exposed in Sections 4.2 and 4.3, namely the likelihood ratio statistic $F_S$, the Fisher statistic $Fi_S$ and the HC statistic $HC^F$. The first two statistics are combined with a deterministic and data-driven model collection, respectively $\mathcal{S}_1$ and $\widehat{\mathcal{S}}_{\text{Lasso}}$, as well as

with a Bonferroni (**B**) or Permutation (**P**) calibration. Note that in practice, we consider the Lars approximation of the Lasso regularization path (Efron et al. 2004), in order to classify variables according to the order in which they enter the activated set of variables and construct $\widehat{\mathcal{S}}_{\text{Lasso}}$. Figure 4.2 summarizes the legend used in the following graphical representations for those seven testing strategies.

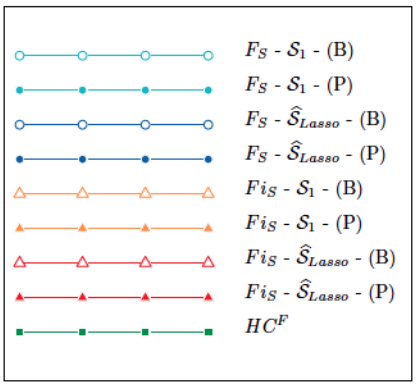



Figure 4.2 – *Colors, symbols and line types used for representing the seven strategies in Figures 4.3, 4.4, 4.5, 4.6*

### Validation of Type I Error Control

**Control Under $\mathcal{H}_{00}$.** Table 4.2 presents level checks under a restricted null hypothesis $\mathcal{H}_{00}$, such that $\beta^{(1)} = \beta^{(2)} = 0$, along with 95% Gaussian confidence intervals. Note that confidence intervals for $n = 100$ are based upon 500 simulations only, and therefore larger than other confidence intervals.

As expected, the Bonferroni calibration combined with the majoration of quantiles or data-driven model collections is, by far, much too conservative. Even with the Fisher statistic, for which we know the exact quantile, it is unthinkable to use Bonferroni calibration as soon as adopt data-driven model collections instead of deterministic ones. Last but not least, the Monte-Carlo calibration works quite well for the two-sample HC statistic.

**Control Under $\mathcal{H}_0$ Only.** Figure 4.3 presents level checks under $\mathcal{H}_0$ but with non null $\beta^{(1)} = \beta^{(2)} \neq 0$. Conclusions are perfectly similar to the case $\mathcal{H}_{00}$: all methods behave well, except the Bonferroni calibration for $F_S$ (using both model collections) and for $Fis_S$ as soon as we use the data-driven model collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ instead of the deterministic collection $\mathcal{S}_1$.

**Power Analysis.** For clarity purposes, we split the results into four different figures. Figure 4.4 represents power performances for the likelihood ratio statistic, Figure 4.5 focuses on the Fisher statistic while Figure 4.6 compares the previous two to HC. Naturally settings 1 and 3 are easier than settings 2 and 4, since there are fewer common coefficients.

(a) $F_S$ statistic

| Model collection | $\mathcal{S}_1$ | | $\widehat{\mathcal{S}}_{Lasso}$ | |
|---|---|---|---|---|
| Calibration | **(B)** | **(P)** | **(B)** | **(P)** |
| $n = 25$ | $0.1 \pm 0.2$ | $6 \pm 1.5$ | $0.1 \pm 0.2$ | $6 \pm 1.5$ |
| $n = 50$ | $0.1 \pm 0.2$ | $4.1 \pm 1.2$ | $0.1 \pm 0.2$ | $4.1 \pm 1.2$ |
| $n = 100^*$ | $0 \pm 0$ | $6.5 \pm 2.2$ | $0 \pm 0$ | $6.5 \pm 2.2$ |

(b) $Fi_S$ statistic

| Model collection | $\mathcal{S}_1$ | | $\widehat{\mathcal{S}}_{Lasso}$ | |
|---|---|---|---|---|
| Calibration | **(B)** | **(P)** | **(B)** | **(P)** |
| $n = 25$ | $5.1 \pm 1.4$ | $5.3 \pm 1.4$ | $0.8 \pm 0.6$ | $5 \pm 1.4$ |
| $n = 50$ | $5 \pm 1.4$ | $5.6 \pm 1.4$ | $0.3 \pm 0.3$ | $4.8 \pm 1.3$ |
| $n = 100^*$ | $5.1 \pm 1.9$ | $4.8 \pm 1.9$ | $0 \pm 0$ | $3.9 \pm 1.7$ |

(c) $HC$ statistic

| Statistic | $HC^F$ |
|---|---|
| $n = 25$ | $5.2 \pm 1.4$ |
| $n = 50$ | $5.5 \pm 1.4$ |
| $n = 100^*$ | $5.5 \pm 2$ |

Table 4.2 – *Estimated test levels in percentage along with 95% Gaussian confidence interval (in percentage) under $\mathcal{H}_{00}$ for the seven different strategies, based upon 1000 simulations. *: simulations for $n = 100$ are based on only 500 simulated samples.*

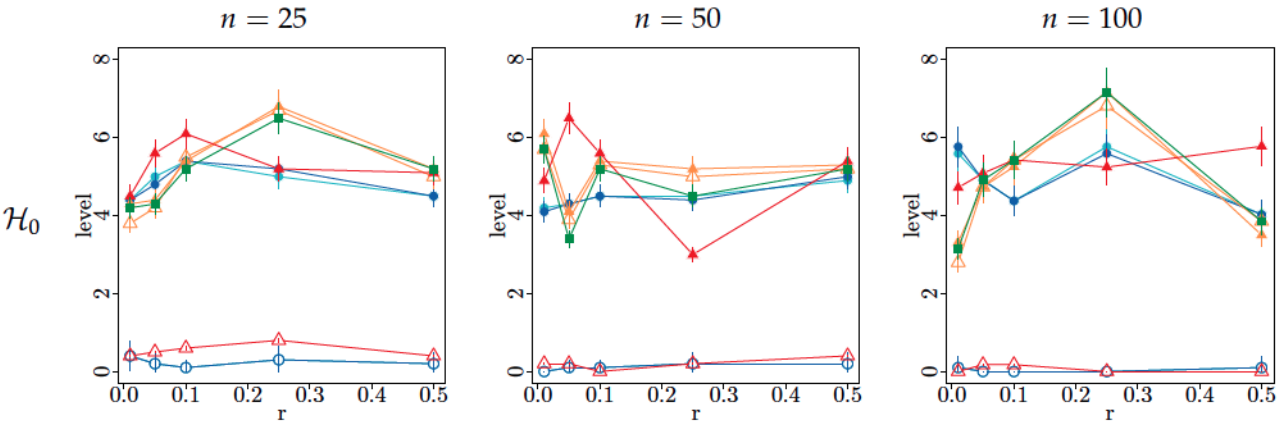



Figure 4.3 – *Estimated test levels in percentage under $\mathcal{H}_0$ for the seven different strategies for varying magnitudes of common non null coefficients, based upon 1000 simulations.*

Focusing on either the usual Fisher statistic $Fi_S$ in Figure 4.5 or the likelihood ratio statistic $F_S$ in Figure 4.4, the Bonferroni calibration is always less powerful than the calibration by permutation, but results are not as bad as we would expect from the level values obtained in Table 4.2 and Figure 4.3 in settings 1 and 3 for $F_S$.

The influence of data-driven model collections is stronger on Fisher statistics, but is always alleviated in settings 1 and 3 where there is never more than one common coefficients. Indeed, under these settings any model of size 1 containing one of the variables activated in only $\beta^{(2)}$ can

suffice to reject the null, which is why collection $\mathcal{S}_1$ performs actually very well. However, in more complex settings 2 and 4, where larger models are required to correct for common effects, model collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ performs a lot better than the collection $\mathcal{S}_1$.

Figure 4.6 compares the previous multivariate statistics to HC. Both the likelihood ratio and Fisher statistics are represented with calibration by permutation and data-driven model collection, which performed best in general. Roughly speaking, the likelihood ratio statistic seems to outperform the usual Fisher statistic in settings 1 and 3, especially when the number of observations is very small ($n = 25$). Yet as soon as the number of common coefficients increases, in settings 2 and 4, the Fisher statistic seems to correct better for common effects. The HC statistic, though really interesting in terms of computing time (compared to the 1000 permutations required for the calibration of $F_S$ and $Fi_S$), would not retain our attention by its performances on this Figure. However, as mentioned in Section 4.3, contrary to other statistics, HC does not take into account the correction for common effects. Therefore, we would like to try and correct for these in a two-step approach, described and evaluated numerically in the next paragraph.

**Two-Step Strategy Correcting for Strong Common Effects.** To correct for strong confounding common effects, we suggest to run the HC test in two-step approach: first, correct for possible common effects by fitting a joint model of reduced dimention, second, apply the HC strategy on the residuals. The best joint subset of covariates $S$ is chosen with the package `LINselect` based upon the procedure developed by Baraud et al. (2010), in order to select the model with minimum Euclidean risk along the Lasso regularization path. We then fit the joint ordinary least square estimator $\widehat{\beta}_S$, and compute $HC^F$ on the residuals $\widetilde{\mathbf{Y}}^i = \mathbf{Y}^{(i)} - \mathbf{X}_S^{(i)}\widehat{\beta}_S$, for $i = 1, 2$. Since at least $F_S$ seems to suffer from the existence of common effects in settings 2 and 4, for comprehensiveness we also try this strategy on $F_S$ and $Fi_S$. The legend associated with the following figures appears in Figure 4.7.

As we have no theoretical developments yet to guarantee the control of type I error for this two-step approach, we rely on Table 4.3 and Figure 4.8 to guarantee that the required 5% level is satisfied. The two-step strategy appears to be deleterious to the Fisher statistic, which becomes much too conservative as soon as the magnitude of common effects increases.

Figure 4.9 compare the test statistics represented in Figure 4.6 to their two-step counterpart. Fortunately, the two-step approach does not impair the results when no common effects should be detected in settings 1, but improves already a little the results of HC in settings 3. As seen from level checks, the two-steps approach damages the performances of the Fisher statistics when common effects are to be detected as in settings 2 and 4. On the contrary, both the likelihood ratio statistic and HC statistic get improved by the two-step approach. In settings 2, the likelihood ratio statistics outperforms the usual Fisher statistic when $n = 25$ and performs as good when $n = 50$ or 100. Finally, both the likelihood ratio and HC statistics, which had a lot of troubles dealing with settings 4, now perform far better, reaching really outstanding powers for $n = 50$ and even outperforming the Fisher statistics.

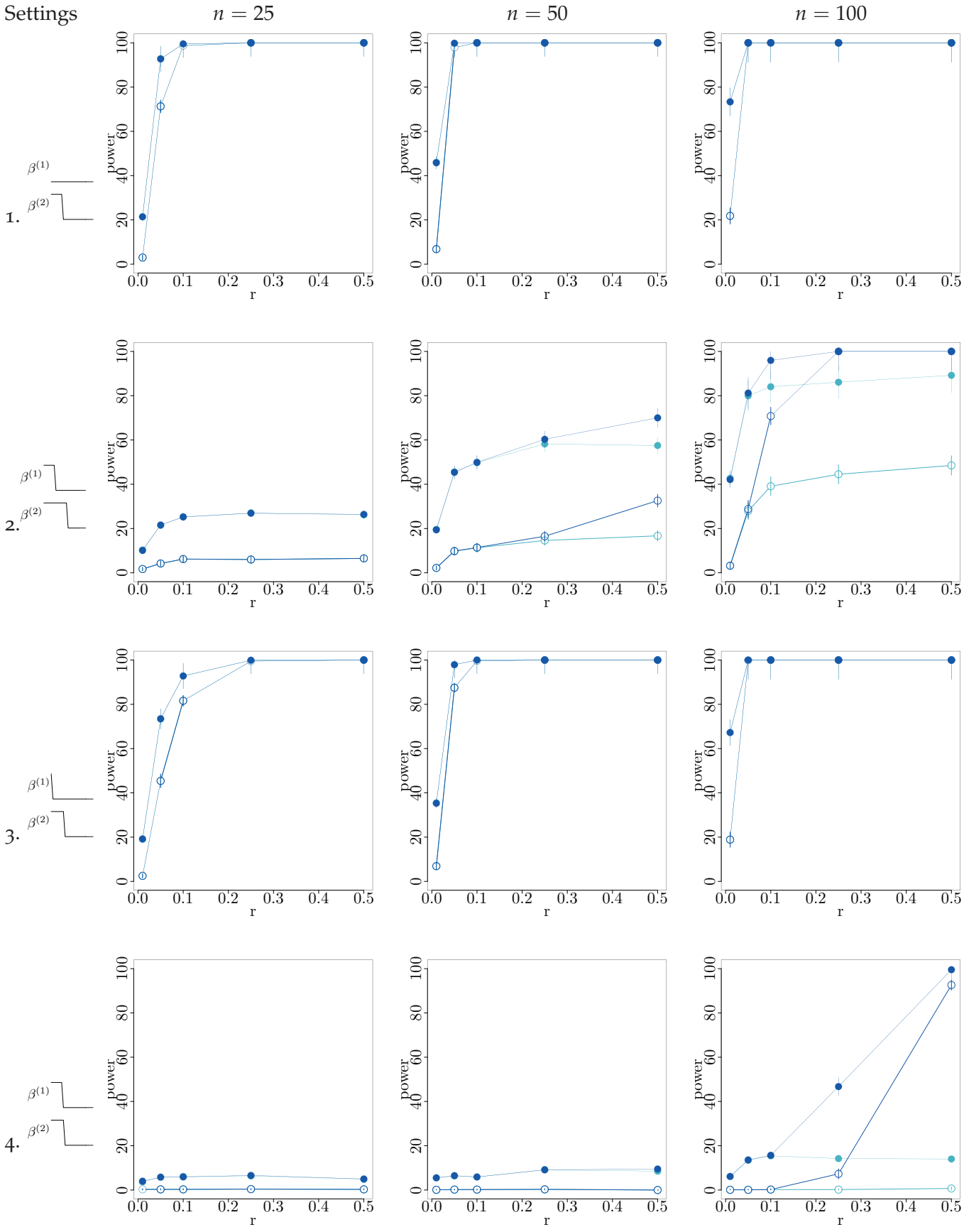



Figure 4.4 – *Power for likelihood ratio statistics, comparing the influence of the choices of model collection and calibration on power under different settings.*

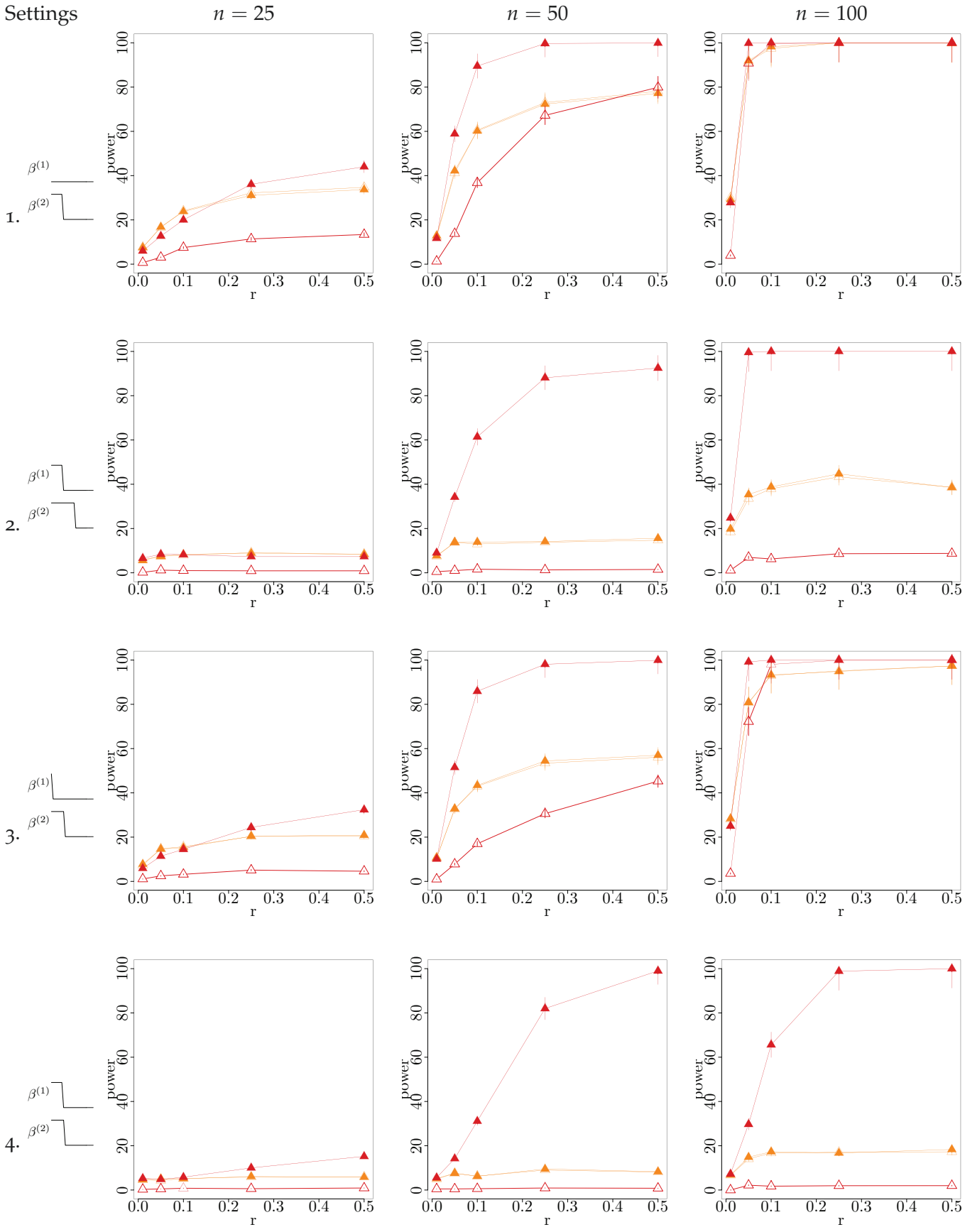



Figure 4.5 – *Power for Fisher statistics, comparing the influence of the choices of model collection and calibration on power under different settings.*

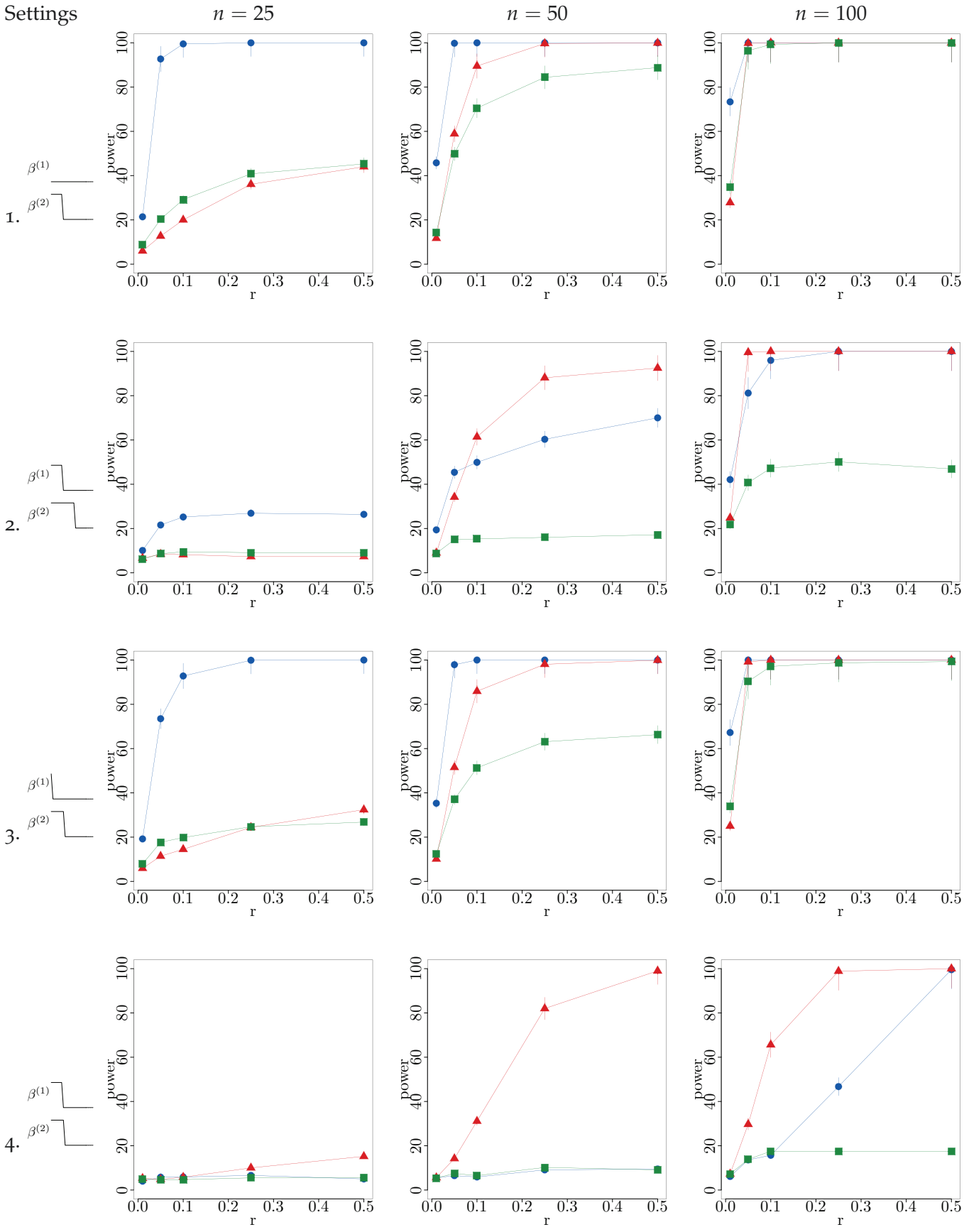



Figure 4.6 – *Power comparisons for likelihood ratio, Fisher and HC statistics.*

(a) $F_S$ statistic

| Model collection | $\mathcal{S}_1$ | $\widehat{\mathcal{S}}_{Lasso}$ |
|---|---|---|
| Calibration | **(P2)** | **(P2)** |
| $n = 25$ | $6.3 \pm 1.5$ | $6.4 \pm 1.5$ |
| $n = 50$ | $4.3 \pm 1.3$ | $4.1 \pm 1.2$ |
| $n = 100^*$ | $6.7 \pm 2.2$ | $6.5 \pm 2.2$ |

(b) $Fi_S$ statistic

| Model collection | $\mathcal{S}_1$ | $\widehat{\mathcal{S}}_{Lasso}$ |
|---|---|---|
| Calibration | **(P2)** | **(P2)** |
| $n = 25$ | $5.3 \pm 1.4$ | $4.8 \pm 1.3$ |
| $n = 50$ | $5.4 \pm 1.4$ | $4.5 \pm 1.3$ |
| $n = 100^*$ | $5.1 \pm 1.9$ | $4.2 \pm 1.8$ |

(c) $HC$ statistic

| Statistic | $HC_2^F$ |
|---|---|
| $n = 25$ | $5.2 \pm 1.4$ |
| $n = 50$ | $5.5 \pm 1.4$ |
| $n = 100^*$ | $5.5 \pm 2$ |

Table 4.3 – *Level checks under $\mathcal{H}_{00}$ for two-step strategies in percentages, along with 95% Gaussian confidence intervals (in percentages). *: simulations for $n = 100$ done for only 500 simulated samples.*

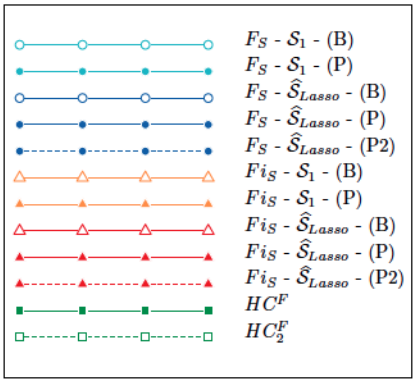



Figure 4.7 – *Colors, symbols and line types used for representing the seven strategies in Figures 4.8 and 4.9.*

### 4.4.2 Real Transcriptomic Data

The procedures developed in Sections 4.2 and 4.3 can be adapted to the case Gaussian graphical models as in Verzelen and Villers (2009). The idea is to run for each gene in the network a neighborhood test conducted at level $\alpha/p$ in order to correct for multiple testing.

At first sight, we would like to test for the equality of neighborhoods, in other words testing for the equality of the supports of $\beta^{(1)}$ and $\beta^{(2)}$. However, in terms of biological interpretation, the test for $\beta^{(1)} = \beta^{(2)}$ also brings a relevant answer to question of whether the regulatory relationships are altered in sample 2 compared to sample 1. As such, we can detect activations replaced by inhibitions, or differences in the strength of the activation or inhibition.

We apply the procedures on the cancer dataset presented in Chapter 3. We run all permuted tests as well as the HC-test in two steps at level $\alpha/62$,

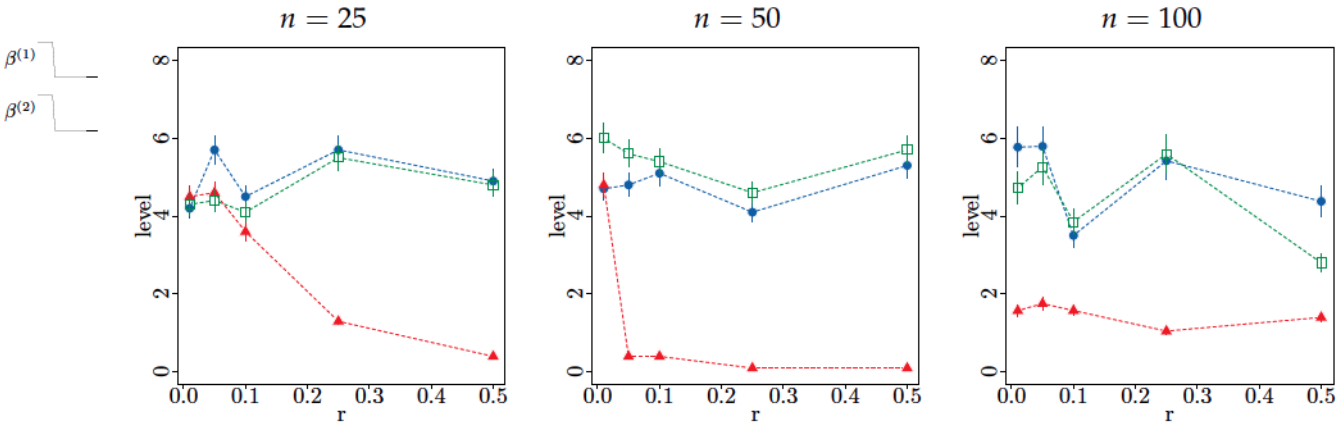



Figure 4.8 – *Estimated test levels in percentages under $\mathcal{H}_0$ for two-step strategies for varying magnitudes of common non null coefficients, based upon 1000 simulations.*

with $\alpha = 5\%$. Neighborhoods rejected at that level for at least one of the statistics are reported on Table 4.4.

| | $HC_2^F$ | $F_S$ | | | $Fi_S$ | | |
|---|---|---|---|---|---|---|---|
| | | $\mathcal{S}_1$ | $\widehat{\mathcal{S}}_{\text{Lasso}}$ | | $\mathcal{S}_1$ | $\widehat{\mathcal{S}}_{\text{Lasso}}$ | |
| | | (P) | (P) | (P2) | (P) | (P) | (P2) |
| PSMB8 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| TAP1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CXCL10 | 0 | 1 | 1 | 0 | 1 | 1 | 1 |
| CXCL9 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| HLA-DOB | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CYBB | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| NCF2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CXCL11 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| CD247 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CD2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| CD38 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| RYR2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| RYR3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

Table 4.4 – *Summary of rejected neighbordhood tests at level 5% corrected by Bonferroni, according to the different testing procedures.*

## 4.5 Discussion

We develop two different testing schemes tackling the problem of two-sample homogeneity tests. We suggest an adaptive likelihood-ratio test which reaches minimax high-dimensional rates of testing, which actually demonstrates great empirical performances thanks to a calibration by permutation which achieves the required type-I error rate. We would like to confirm those performances under more complex simulated designs.

We note that the calibration by permutation is highly time-consuming, which can be highly restrictive in the range of possible applications, par-

ticularly if we think of Gaussian graphical models applied to the inference of gene regulatory networks.

On the contrary, the two-step adaptation of higher-criticism looks rather promising for an excellent performance over computing-time ratio, as observed empirically. In the view of its interesting empirical performances, we would like to explore the theoretical properties of the two-step higher-criticism approach.

## 4.6 Technical Details

This Section explicits the upper bounds $\widetilde{Q}_{2,|S|}(u|\mathbf{X}_S)$ and $\widetilde{Q}_{3,|S|}(u|\mathbf{X}_S)$. Because of the symmetry between $F_{2,S}$ and $F_{3,S}$, we only provide developments for $F_{S,2}$. Let us note $a = (a_1, \ldots, a_{|S|})$ the positive eigenvalues of

$$\frac{n_1}{n_2(n_1-|S|)}\mathbf{X}_S^{(2)}\left[(\mathbf{X}_S^{(1)\intercal}\mathbf{X}_S^{(1)})^{-1} + (\mathbf{X}_S^{(2)\intercal}\mathbf{X}_S^{(2)})^{-1}\right]\mathbf{X}_S^{(2)\intercal}\,.$$

**Definition 4.1** (Recall of the definition of the upper-bound $\widetilde{Q}_{2,|S|}(u|\mathbf{X}_S)$) *Consider some number $u > \|a\|_1$. If all the components of $a$ are equal, then we take*

$$\lambda^* = \frac{u - \|a\|_1}{2u(\|a\|_\infty + \frac{\|a\|_1}{n_1-|S|})}$$

*If $a$ is not a constant vector, then we define $\lambda^*$ by*

$$\begin{aligned} b &:= \frac{\|a\|_1 u}{\|a\|_\infty(n_1-|S|)} + u + \frac{\|a\|_2^2}{\|a\|_\infty} - \|a\|_1, \\ \Delta &:= b^2 - \frac{4u\,(u-\|a\|_1)}{(n_1-|S|)\|a\|_\infty}\left(\|a\|_1 - \frac{\|a\|_2^2}{\|a\|_\infty}\right), \qquad (4.20) \\ \lambda^* &:= \frac{1}{\frac{4u}{n_1-\|S\|}\left(\|a\|_1 - \frac{\|a\|_2^2}{\|a\|_\infty}\right)}\left(b - \sqrt{\Delta}\right) \qquad (4.21) \end{aligned}$$

*We recall that $\widetilde{Q}_{2,|S|}(u|\mathbf{X}_S)$ is defined as follows*

$$\widetilde{Q}_{2,|S|}(u|\mathbf{X}_S) := \exp\left[-\frac{1}{2}\sum_{i=1}^{|S|}\log\left(1-2\lambda^* a_i\right) - \frac{n_1-|S|}{2}\log\left(1+\frac{2\lambda^* u}{n_1-|S|}\right)\right].$$

*Proof of Proposition 4.3.* For the sake of simplicity, we note $N = n_1 - |S|$, $(Z_1, \ldots, Z_{|S|})$ a standard Gaussian random vector and $W_N$ a $\chi^2$ random variable with $N$ degrees of freedom. We apply Laplace method to upper bound $\mathbb{P}[F_{2,S} \geq u]$:

$$\begin{aligned} \mathbb{P}[F_{2,S} \geq u] &= \mathbb{P}\left[\sum_{i=1}^{|S|} a_i Z_i^2 \geq u W_N/N\right] \leq \inf_{\lambda>0}\mathbb{E}\exp\left[\lambda\sum_{i=1}^{|S|} a_i Z_i^2 - \lambda u W_N/N\right] \\ &\leq \inf_{0<\lambda<|a|_\infty/2}\exp\left[\psi_u(\lambda)\right], \end{aligned}$$

where

$$\psi_u(\lambda) = -\frac{1}{2}\sum_{i=1}^{|S|}\log(1-2\lambda a_i) - \frac{N}{2}\log\left(1+\frac{2\lambda u}{N}\right).$$

The sharpest upper-bound is given by the value $\lambda^*$ which minimizes $\psi_u(\lambda)$. We obtain an approximation of $\lambda^*$ by cancelling the second-order approximation of its derivative. Deriving $\psi_u$ gives

$$\psi_u'(\lambda) = \sum_{i=1}^{|S|}\frac{a_i}{1-2\lambda a_i} - \frac{u}{1+\frac{2\lambda u}{N}},$$

which admits the following second order approximation :

$$\|a\|_1 + \frac{2\lambda\|a\|_2^2}{1-2\|a\|_\infty\lambda} - \frac{u}{1+\frac{2\lambda u}{N}} \,. \tag{4.22}$$

Cancelling this quantity amounts to solving a polynomial equation of the second degree. The smallest solution of this equation leads to the desired $\lambda^*$. □

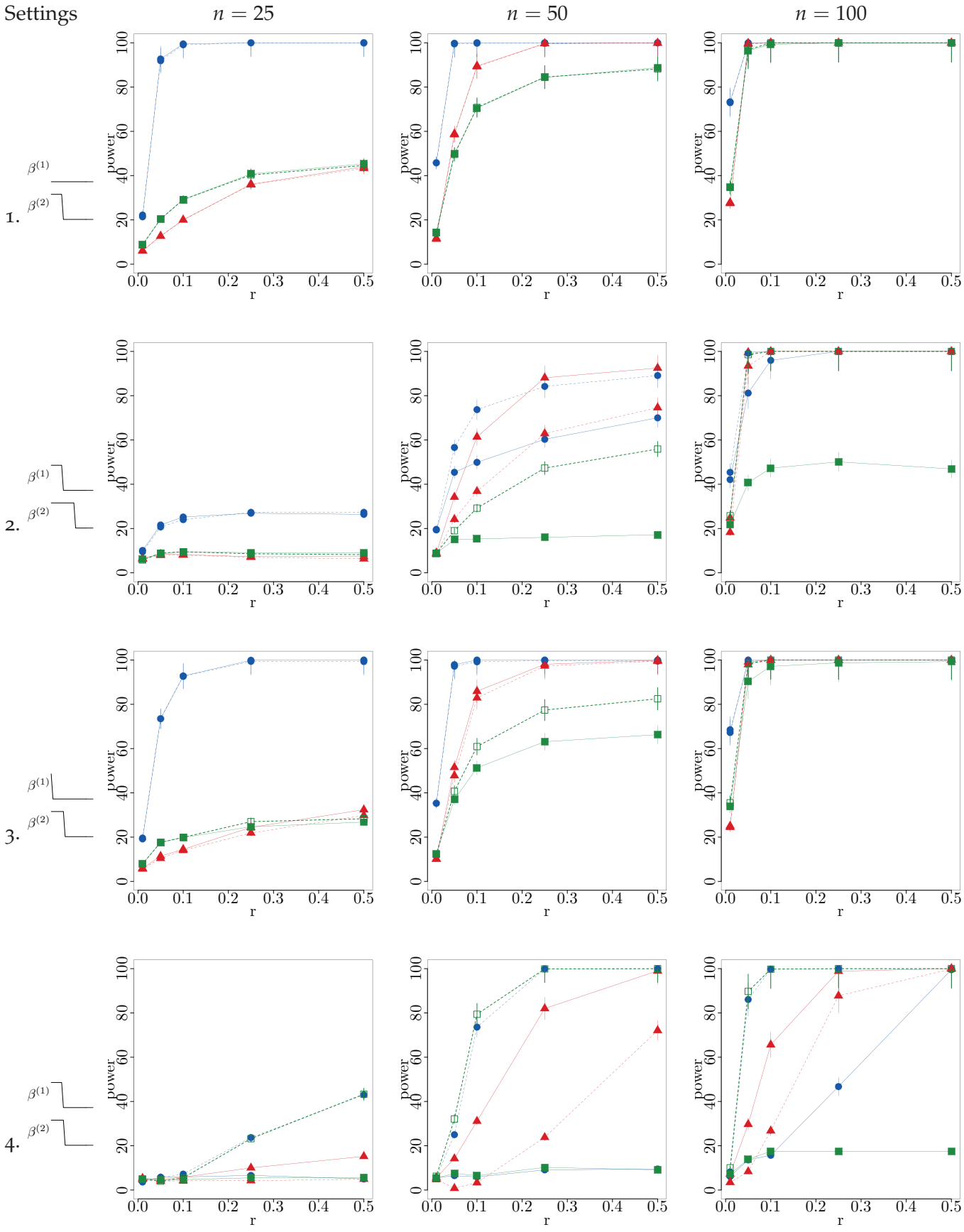



Figure 4.9 – *Influence of the correction for common effects on the power performances of the likelihood ratio, Fisher and HC statistics.*

# Discussion and Perspectives

This thesis investigates the inference of high-dimensional Gaussian graphical models from non identically and independently distributed transciptomic data in the objective of recovering gene regulatory networks. In the context of high-dimensional statistics, the heterogeneity of the dataset fruitfully paves the way to the definition of structured regularizers via weighted and block-sparse penalties. We also examine the crucial issue of validating the answers provided by high-dimensional estimators.

Admittedly, the application of Gaussian graphical models to real transcriptomic dataset reveals the limitations of our modeling of regulatory phenomena. As exemplified by *E. coli* S.O.S. network in Chapter 2, the multiplicity of actors and levels of regulation left out by transcriptomic data compromises the interpretation of inferred networks. Yet, it is hard to imagine a solid statistical model integrating data from all relevant fields of data (proteomic, transcriptomic, genetic, methylation, etc).

Besides, it remains difficult to correctly evaluate the performances of our methods, even to define what a correct evaluation would be. Indeed, simulated experiments provide a comparison of competing estimators in perfectly identified and controlled settings. Yet they are too close to our statistical modeling assumptions to provide a realistic evaluation of actual performances on real datasets. On the contrary, applications to real datasets lack of trustful benchmarks: even for model species like *E. coli*, it is not obvious to reconstruct the actual set of regulations that should take place in a given condition. Available gene regulatory networks might still miss some actual regulations or include some regulations that exclusively happen under some particular stress but not in the conditions under study. Those incertainties result in fuzzy estimations of false negatives and false positives.

As a result, we are avidly waiting for the emergence of clear benchmark networks validated experimentally on small model species in order to finally evaluate to what extent Gaussian graphical models actually capture the transcriptomic regulatory mechanisms in place.

As a by-product, there are some information theoretic questions still pending as to how difficult the question of infering Gaussian graphical models really is, depending on the actual structure of biological networks. Information theoretic limits in high-dimensional linear regression state that depending on the number of observations available and the number of variables considered, one cannot hope to recover more than a certain number edges (Wainwright 2009b, Verzelen 2012).

Information theoretic results about model selection in Gaussian graphical models underline the difficulty associated, again, with a growing

number of neighbors or the detection of too small conditional dependencies (Wang et al. 2010), but most importantly difficulties arising with large eigenvalues of the partial correlation matrix (Anandkumar et al. 2011). In other words, star-shaped networks where some few genes play the role of hubs or highly correlated subsets of genes, both highly probable scenarios, could be troublesome to the statistical inference, even under the assumption that the data perfectly follows some unknown Gaussian graphical model.

These results could explain why in practice we often observe that the inferred networks are highly unstable from a strict point of view, while competing paths in fact roughly speaking reflect the same flow of information: a path from $i$ to $j$ being replaced by a path from $i$ to $k$ to $j$. This phenomenom leaves the feeling that considering the design proportions available or the high-levels of correlation among genes, we might be too ambitious looking for graphical representations at a so fine level. In other words, it could be beneficial to look for less precise representations but where we could hope to obtain more robust results.

In that spirit, causal inference based upon intervention data sets (Maathuis et al. 2010, Hauser and Bühlmann 2012, Bühlmann 2012a) is a way to provide effective answers to biologists while zooming out of the problem of dissecting regulatory mechanisms by integrating out the uncertainties about the precise chain of regulations.

Another way to change the scale of analysis would be to build hierarchical Gaussian graphical models, considering too highly correlated genes as repeated measurements of a single *metagene*. The model would ignore to recover the conditional dependences among these highly correlatd subsets genes but focus instead on conditional dependences among metagenes. We could hope that combining redondant information about the same main regulations would add robustness to the inference process. We are currently investigating to what extent some $\ell_{1,\infty}$ block-regularization could solve such hierarchical models.

# Appendix A

## A.1 Proofs for Chapter 3

### A.1.1 Hölder inequalities for cooperative norms (Proposition 3.2)

Consider $x$ and $y \in \mathbb{R}^p$. By discarding successively the negative terms, the scalar product can be upper bounded by:

$$\begin{aligned}\langle x, y\rangle &= \langle x^+, y\rangle + \langle x^-, y\rangle \\ &\leq \langle x^+, y^+\rangle + \langle x^-, y^-\rangle \\ &\leq \langle x^+, y^+\rangle + \langle x^-, y^-\rangle,\end{aligned}$$

such that

$$|\langle x, y\rangle| \leq |\langle x^+, y^+\rangle| + |\langle x^-, y^-\rangle|.$$

Let us now apply Hölder's inequality for mixed norms on each of the two terms:

$$\begin{aligned}|\langle x, y\rangle| &\leq \|x^+\|_{p,q}\|y^+\|_{p',q'} + \|x^-\|_{p,a}\|y^-\|_{p',q'} \\ &\leq \|x\|_{\text{coop},p,q}\|y\|_{\text{coop},p',q'}.\end{aligned}$$

To prove that $\|\cdot\|_{\text{coop},p,q}$ actually is the dual norm of $\|\cdot\|_{\text{coop},p',q'}$, we need to exhibit $x$ and $y$ such that $\|x\|_{\text{coop},p,q} \leq 1$ and $< x, y >= \|y\|_{\text{coop},p',q'}$.

### A.1.2 Optimality conditions for the coop-Lasso (Theorem 3.4)

Plug the characterization of the subdifferential of a norm (3.8) in combination with the definition of the coop dual norm into Equation (3.5) to obtain:

$$\begin{cases}\max_{k=1,\dots,K}(\|z^+_{\mathcal{G}_k}\|_2, \|z^-_{\mathcal{G}_k}\|) \leq 1 & \text{if } \hat{\boldsymbol{\beta}} = 0 \\ \max_{k=1,\dots,K}(\|z^+_{\mathcal{G}_k}\|_2, \|z^-_{\mathcal{G}_k}\|) = 1 \text{ and } \langle \hat{\boldsymbol{\beta}}, z\rangle = \|\hat{\boldsymbol{\beta}}\|_{\text{coop}} & \text{otherwise.}\end{cases} \tag{A.1}$$

Now, recall that by Hölder's inequality, $|\langle \hat{\boldsymbol{\beta}}, z\rangle| \leq \|\hat{\boldsymbol{\beta}}\|_{\text{coop}}\|z\|_{\text{coop}^\star}$. Therefore, the equality in (A.1) is only possible if for every group $\mathcal{G}_k$ such that $\hat{\boldsymbol{\beta}}_{\mathcal{G}_k} \neq 0$, Hölder's inequality is saturated and $\|z_{\mathcal{G}_k}\|_{\text{coop}^\star} = 1$. Note that since the coop-norm dissociates the positive and negative parts of $\hat{\boldsymbol{\beta}}$, the second equality constraint is only active on signed subgroups, but the first equality adds a companion constraint on complementary indices. For clarity, we introduce a notation for signed subgroups $s_r$, $r = 1, \dots, 2K$, defined such that for every $k = 1, \dots, K$:

$$s_{2k-1} = \{j \in \mathcal{G}_k, \hat{\boldsymbol{\beta}}_j > 0\} \quad \text{and} \quad s_{2k} = \{j \in \mathcal{G}_k, \hat{\boldsymbol{\beta}}_j < 0\}.$$

In this case, the second equality constraint implies for every activated signed subgroup $s_k$ : $\langle \hat{\boldsymbol{\beta}}_{s_k}, z_{s_k}\rangle = \|\hat{\boldsymbol{\beta}}_{s_k}\|_{\text{coop}} = \|\hat{\boldsymbol{\beta}}_{s_k}\|$, which leads to the required expression

$$z_{s_k} = \frac{\hat{\boldsymbol{\beta}}_{s_k}}{\|\hat{\boldsymbol{\beta}}_{s_k}\|}.$$

To guarantee in turn the first equality constraint, the subdifferential requires that for every activated sign subgroup $s_k$, coefficients from a different sign or equal to zero should correspond to a subgradient with opposite sign:

$$\text{sign}(z_{\mathcal{G}_k \setminus s_k}) = -\text{sign}(z_{s_k}). \tag{A.2}$$

Since $\|z_{s_k}\|_{\text{coop}^\star} = 1$, it suffices to require $\|z_{\mathcal{G}_k}\|_{\text{coop}^\star} \leq 1$ to obtain (A.2).

### A.1.3 Support Recovery (Theorem 3.6)

We follow the three-step proof technique proposed by Yuan and Lin (2007b) for the Lasso, applied by Bach (2008b) for the group-Lasso, and referred to as *primal-dual witness construction* proofs in Negahban and Wainwright (2011), Obozinski et al. (2011), Wainwright (2009a), Jalali et al. (2011):

1. restrict the estimation problem to the true support and complete this estimate by 0 outside the true support, thereby defining a primal solution with desired support;
2. exhibit the subgradient (or dual solution, as explained in Section 3.2.3) associated with the primal solution of step 1;
3. prove that this primal-dual pair is optimal asymptotically.

The main differences between asymptotic results as in Yuan and Lin (2007b) or Bach (2008b) and non-asymptotic results as in Negahban and Wainwright (2011), Obozinski et al. (2011), Wainwright (2009a), Jalali et al. (2011) lie in step 3: this last line of work gets rid of the asymptotic control on error terms by restricting themselves to events of high-probability on which error terms are controlled.

To end the proof, remark that under assumption (A2), the solution is unique, leading to the conclusion that the coop-Lasso estimator equal to this artificial estimate with probability tending to 1.

As a first step, we prove two simple lemmas. Lemma A.1 states that the coop-Lasso estimate, restricted on the true support $\mathcal{S}$, is consistent when $\lambda_n \to 0$. Lemma A.2 provides the basis for the inequalities (3.10) and (3.11) that express our irrepresentable conditions.

**Lemma A.1** *Assuming (A1-3), let $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ be the unique minimizer of the regression problem restricted to the true support $\mathcal{S}$:*

$$\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} = \arg\min_{\mathbf{v}\in\mathbb{R}^{|\mathcal{S}|}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}_{.\mathcal{S}}\mathbf{v}\|_n^2 + \lambda_n \sum_{k:\mathcal{S}_k\neq\emptyset} w_k(\|\mathbf{v}^+_{\mathcal{S}_k}\| + \|\mathbf{v}^-_{\mathcal{S}_k}\|) \ ,$$

*where $\|\cdot\|_n = \|\cdot\|/n$ denotes the empirical norm.*

*If $\lambda_n \to 0$, then $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} \xrightarrow{P} \boldsymbol{\beta}^\star_{\mathcal{S}}$.*

*Proof.* This lemma stems from standard results of M-estimation (Van der Vaart 1998). Let $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^\star$, and write $\boldsymbol{\Psi}^n = \mathbf{X}^\intercal\mathbf{X}/n$. If $\lambda_n \to 0$, then under (A1-2), for any $\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|}$

$$\begin{aligned} Z_n(\mathbf{v}) &= \frac{1}{2}\|\mathbf{y} - \mathbf{X}_{.\mathcal{S}}\mathbf{v}\|_n^2 + \lambda_n \sum_{k:\mathcal{S}_k\neq\emptyset} w_k(\|\mathbf{v}^+_{\mathcal{S}_k}\| + \|\mathbf{v}^-_{\mathcal{S}_k}\|) \\ &= \frac{1}{2}(\boldsymbol{\beta}^\star_{\mathcal{S}} - \mathbf{v})^\intercal\boldsymbol{\Psi}^n_{\mathcal{S}\mathcal{S}}(\boldsymbol{\beta}^\star_{\mathcal{S}} - \mathbf{v}) - \frac{1}{n}\boldsymbol{\varepsilon}^\intercal\mathbf{X}_{.\mathcal{S}}(\boldsymbol{\beta}^\star_{\mathcal{S}} - \mathbf{v}) + \frac{\boldsymbol{\varepsilon}^\intercal\boldsymbol{\varepsilon}}{2n} \\ &\qquad + \lambda_n \sum_{k,\mathcal{S}_k\neq\emptyset} w_k(\|\mathbf{v}^+_{\mathcal{S}_k}\| + \|\mathbf{v}^-_{\mathcal{S}_k}\|) \end{aligned}$$

tends in probability to

$$Z(\mathbf{v}) = \frac{1}{2}(\boldsymbol{\beta}^\star_{\mathcal{S}} - \mathbf{v})^\intercal\boldsymbol{\Psi}_{\mathcal{S}\mathcal{S}}(\boldsymbol{\beta}^\star_{\mathcal{S}} - \mathbf{v}) + \frac{1}{2}\sigma^2.$$

It follows from the strict convexity of $Z_n$ that $\arg\min Z_n(\mathbf{v}) \xrightarrow{P} \arg\min Z(\mathbf{v}) = \boldsymbol{\beta}^\star_{\mathcal{S}}$ (Knight and Fu 2000), which ends the proof. □

**Lemma A.2** *Consider a sequence of random variables $S_n$ such that $S_n \xrightarrow{P} S$. Suppose there exists $\delta > 0$ such that for a given norm $\mu$ the limit $S$ is bounded away from 1:*

$$\mu(S) \leq 1 - \delta \ .$$

*Then,*

$$\mathbb{P}(\mu(S_n) \leq 1) \to 1 \ .$$

*Proof.* By triangular inequality and thanks to the constraint on $\mu(S)$:

$$\mathbb{P}(\mu(S_n) \leq 1) \geq \mathbb{P}\big(\mu(S_n - S) \leq 1 - \mu(S)\big) \geq \mathbb{P}(\mu(S_n - S) \leq \delta) \ ,$$

Convergence in probability of $S_n$ to $S$ concludes the proof:

$$\mathbb{P}(\mu(S_n - S) \leq \delta) \to 1 \ , \quad \text{therefore} \quad \mathbb{P}(\mu(S_n) \leq 1) \to 1 \ .$$

□

Let us consider the full vector $\tilde{\boldsymbol{\beta}}^n$ with coefficients $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ defined as in Lemma A.1 and other coefficients null, $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}^c} = \mathbf{0}$. We now proceed to the last step of the proof of Theorem 3.6, by proving that $\tilde{\boldsymbol{\beta}}^n$ satisfies the coop-Lasso optimality conditions with probability tending to 1 under the additional conditions (A4-5). The final conclusion then results from the uniqueness of the coop-Lasso estimator.

First, consider optimality conditions with respect to $\boldsymbol{\beta}_{\mathcal{S}}$. As a result of Lemma A.1, the probability that $\tilde{\boldsymbol{\beta}}^n_j \neq 0$ for every $j \in \mathcal{S}$ tends to 1. Thereby, $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ satisfies the conditions of Theorem 3.4 on the restriction of $\mathbf{X}$ to covariates in $\mathcal{S}$ with probability tending to 1. As $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}^c} = \mathbf{0}$, then $\mathbf{X}\tilde{\boldsymbol{\beta}}^n = \mathbf{X}_{\cdot\mathcal{S}}\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ and for every $j \in \mathcal{S}$, $\|\boldsymbol{\varphi}_j(\tilde{\boldsymbol{\beta}}_{\mathcal{S}_k})^n\| = \|\boldsymbol{\varphi}_j(\tilde{\boldsymbol{\beta}}^n_{\mathcal{G}_k})\|$, therefore $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ satisfies optimality conditions of Theorem 3.4 in the original problem with probability tending to 1.

Second, $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}^c}$ should also verify the optimality conditions with probability tending to 1. With assumption (A3), we only have to consider two cases that read:

- if group $k$ is excluded from the support, one must have

$$\mathbb{P}\left(\max\left(\|((\mathbf{X}_{\cdot\mathcal{S}^c_k})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n - \mathbf{y}))^+\|_n, \|((\mathbf{X}_{\cdot\mathcal{S}^c_k})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n - \mathbf{y}))^-\|_n\right) \leq \lambda_n w_k\right) \to 1 \ ; \tag{A.3}$$

- if group $k$ intersects the support, with either positive ($\nu_k = 1$) or negative ($\nu_k = -1$) coefficients, one must have

$$\mathbb{P}\left(\{\nu_k(\mathbf{X}_{\cdot\mathcal{S}^c_k})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n - \mathbf{y}) \succeq \mathbf{0}\} \cap \{\|(\mathbf{X}_{\cdot\mathcal{S}^c_k})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n - \mathbf{y})\|_n \leq \lambda_n w_k\}\right) \to 1 \ . \tag{A.4}$$

To prove (A.3) and (A.4), we study the asymptotics of $(\mathbf{X}_{\cdot\mathcal{S}^c_k})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n - \mathbf{y})/n$ for any group such that $\mathcal{S}^c_k$ is not empty. As a consequence of the existence

of the fourth order moments of the centered random variables $X$ and $Y$, the multivariate central limit theorem applies, yielding:

$$\frac{\mathbf{X}^\intercal\mathbf{X}}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\intercal\mathbf{x}_i = \mathbf{\Psi} + O_P(n^{-1/2}), \qquad \frac{\mathbf{X}^\intercal\boldsymbol{\varepsilon}}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\boldsymbol{\varepsilon}_i = O_P(n^{-1/2}) \tag{A.5}$$

Then, we derive from (A.5) and the definition of $\tilde{\boldsymbol{\beta}}^n$ that

$$\begin{aligned}\frac{1}{n}(\mathbf{X}_{\cdot\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n - \mathbf{y}) &= \frac{1}{n}(\mathbf{X}_{\cdot\mathcal{S}_k^c})^\intercal\mathbf{X}(\tilde{\boldsymbol{\beta}}^n - \boldsymbol{\beta}^\star) - \frac{1}{n}(\mathbf{X}_{\cdot\mathcal{S}_k^c})^\intercal\boldsymbol{\varepsilon} \\ &= \frac{1}{n}(\mathbf{X}_{\cdot\mathcal{S}_k^c})^\intercal\mathbf{X}_{\cdot\mathcal{S}}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} - \boldsymbol{\beta}^\star_{\mathcal{S}}) + O_P(n^{-1/2}) \\ &= \mathbf{\Psi}_{\mathcal{S}_k^c\mathcal{S}}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} - \boldsymbol{\beta}^\star_{\mathcal{S}}) + O_P(n^{-1/2}) \ . \end{aligned} \tag{A.6}$$

While the combination of (A.5) and optimality conditions 3.4 on $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ leads to:

$$\mathbf{\Psi}_{\mathcal{S}\mathcal{S}}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} - \boldsymbol{\beta}^\star_{\mathcal{S}}) = -\lambda_n\mathbf{D}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}})\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} + O_P(n^{-1/2}) \ , \tag{A.7}$$

where $\mathbf{D}(\cdot)$ is the weighting matrix (3.9). Put (A.6) and (A.7) together to finally obtain:

$$\frac{1}{n}(\mathbf{X}_{\cdot\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n - \mathbf{y}) = -\lambda_n\mathbf{\Psi}_{\mathcal{S}_k^c\mathcal{S}}\mathbf{\Psi}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{D}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}})\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}} + O_P(n^{-1/2}) \ . \tag{A.8}$$

Now, define for any $k$ such that $\mathcal{S}_k^c$ is not empty:

$$R_{k,n} = \frac{1}{w_k\lambda_n}\frac{1}{n}(\mathbf{X}_{\cdot\mathcal{S}_k^c})^\intercal(\mathbf{X}\tilde{\boldsymbol{\beta}}^n - \mathbf{y}) \quad \text{and} \quad R_k = -\frac{1}{w_k}\mathbf{\Psi}_{\mathcal{S}_k^c\mathcal{S}}\mathbf{\Psi}_{\mathcal{S}\mathcal{S}}^{-1}\mathbf{D}(\boldsymbol{\beta}^\star_{\mathcal{S}})\boldsymbol{\beta}^\star_{\mathcal{S}} \ ,$$

Limits (A.3) and (A.4) are expressed:

- if group $k$ is excluded from the support, one must have

$$\mathbb{P}\left(\max\left(\|R_{k,n}^+\|, \|R_{k,n}^-\|\right) \leq 1\right) \to 1 \ ;$$

- if group $k$ intersects the support, with either positive ($\nu_k = 1$) or negative ($\nu_k = -1$) coefficients, one must have

$$\mathbb{P}\left(\{\nu_k R_{k,n} \succeq \mathbf{0}\} \cap \{\|(\nu_k R_{k,n})^+\| \leq 1\}\right) \to 1 \ .$$

Remark that, as a continuous function of $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$, $\mathbf{D}(\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}})\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}}$ converges in probability to $\mathbf{D}(\boldsymbol{\beta}^\star_{\mathcal{S}})\boldsymbol{\beta}^\star_{\mathcal{S}}$. Therefore, with a decrease rate for $\lambda_n$ chosen such that $n^{1/2}\lambda_n \to \infty$, equation (A.8) implies

$$R_{k,n} \xrightarrow{P} R_k \ . \tag{A.9}$$

It now suffices to successively apply Lemma A.2 to the appropriate vectors and norms to show that $\tilde{\boldsymbol{\beta}}^n_{\mathcal{S}^c}$ satisfies (A.3) and (A.4):

- if group $k$ is excluded from the support, (A4) assumes that there exists $\eta > 0$, such that

$$\max(\|R_k^+\|, \|R_k^-\|) \leq 1 - \eta \ ,$$

and Lemma A.2 applied to $\mu(u) = \max(\|u^+\|, \|u^-\|)$ provides

$$\mathbb{P}\{\max(\|R_{k,n}^+\|, \|R_{k,n}^-\|) \leq 1\} \to 1 \ .$$

- if group $k$ intersects the support, with either positive ($\nu_k = 1$) or negative ($\nu_k = -1$) coefficients,

$$\begin{aligned}\mathbb{P}\left(\{\|(\nu_k R_{k,n})^+\| \leq 1\} \cap \{\nu_k R_{k,n} \succeq \mathbf{0}\}\right) \\ &= 1 - \mathbb{P}\left(\{\|(\nu_k R_{k,n})^+\| > 1\} \cup \{\nu_k R_{k,n} \prec \mathbf{0}\}\right) \\ &\geq 1 - \mathbb{P}\left(\|(\nu_k R_{k,n})^+\| > 1\right) - \mathbb{P}\left(\nu_k R_{k,n} \prec \mathbf{0}\right) \\ &\geq 1 - \mathbb{P}\left(\max(\|R_{k,n}^+\|, \|R_{k,n}^-\|) > 1\right) - \mathbb{P}\left(\nu_k R_{k,n} \prec \mathbf{0}\right) \quad .\end{aligned}$$

  As previously, the first probability in the sum tends to 0 because of (A4) and Lemma A.2. The second probability tends to 0 from (A5) and of the convergence in probability of $R_{k,n}$ to $R_k$. Therefore the overall probability tends to 1.

Denote by $A_{k,n}$ these events on which coefficients in $\mathcal{S}_k^c$ are set to 0. We just showed that individually for each group $k$ with true null coefficients, $P(A_{k,n}) \to 1$. This implies that,

$$\mathbb{P}\left(\bigcup_{k:\mathcal{S}_k^c \neq \emptyset} A_{k,n}^c\right) \leq \sum_{k:\mathcal{S}_k^c \neq \emptyset} \mathbb{P}\left(A_{k,n}^c\right) \to 0,$$

which in turn concludes the proof:

$$\mathbb{P}\left(\bigcap_{k:\mathcal{S}_k^c \neq \emptyset} A_{k,n}\right) \to 1.$$

□

### A.1.4 Oracle Inequalities stated in Theorem 3.7 and Corollary 3.8

The sketch of the proof is very similar to the one adopted for the Lasso Bickel et al. (2009) or group-Lasso Lounici et al. (2009). Particularly interesting improvements have more recently been suggested in Lounici et al. (2011) and S. Negahban and Yu (2012). Those results make use of dual bounds and equivalence relationships between coop-norms, as exhibited in Section 3.1.

**Lemma A.3** *For every $\boldsymbol{\beta} \in \mathbb{R}^p$, the coop-Lasso solution $\hat{\boldsymbol{\beta}}^{\text{coop}}$ satisfies the following inequalities on the event that $\{\|\mathbf{X}^\intercal \varepsilon / n\|_{coop,\infty} \leq \lambda_n / 2\}$:*

$$\begin{aligned}\|\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n^2 + \lambda_n \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{coop} &\leq \|\mathbf{X}(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\|_n^2 + 4\lambda_n \|\boldsymbol{\beta}_{\mathcal{S}(\boldsymbol{\beta})} - \hat{\boldsymbol{\beta}}^{\text{coop}}_{\mathcal{S}(\boldsymbol{\beta})}\|_{coop}, \\ \|\mathbf{X}^\intercal \mathbf{X}(\hat{\boldsymbol{\beta}}^{\text{coop}} - \boldsymbol{\beta}^\star)/n\|_{coop,\infty} &\leq \frac{3}{2}\lambda_n.\end{aligned}$$

**Proof of lemma A.3**

**First inequality** It follows from the definition of the coop-Lasso given in Equation that for every $\boldsymbol{\beta} \in \mathbb{R}^p$ :

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{coop}}\|_n^2 + 2\lambda_n \|\hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_n^2 + 2\lambda_n \|\boldsymbol{\beta}\|_{\text{coop}}.$$

From the decomposition of $\mathbf{y}$ into $\mathbf{X}\boldsymbol{\beta}^\star + \varepsilon$ we deduce:

$$\begin{aligned}&\|\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n^2 + 2\lambda_n \|\hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} \\ &\leq \|\mathbf{X}(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\|_n^2 + 2\lambda_n \|\boldsymbol{\beta}\|_{\text{coop}} + \frac{2}{n}\varepsilon^\intercal \mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{coop}}). \end{aligned} \tag{A.10}$$

Now, the dual bound derived for the coop-norm provides a bound on the scalar product on the right hand side of inequality (A.10).

$$\frac{2}{n}\varepsilon^\intercal \mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{coop}}) \leq 2\|\frac{\mathbf{X}^\intercal \varepsilon}{n}\|_{\text{coop},\infty} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}}.$$

On event $\mathcal{A} = \{2\|\mathbf{X}^\intercal \varepsilon / n\|_{\text{coop},\infty} \leq \lambda_n\}$, inequality (A.10) rewrites

$$\begin{aligned}&\|\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n^2 + \lambda_n \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} \\ &\quad \leq \|\mathbf{X}(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\|_n^2 + 2\lambda_n \left( \|\boldsymbol{\beta}\|_{\text{coop}} - \|\hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} + \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} \right)\end{aligned}$$

A close look at the last term on the right hand side shows that all terms corresponding groups $\mathcal{G}_k$ such that $\|\boldsymbol{\beta}_{\mathcal{G}_k}\|_{\text{coop}} = 0$ disappear. Denoting by $\mathcal{S}(\boldsymbol{\beta})$ the group support of $\boldsymbol{\beta}$, that is to say $\{k = 1 \dots, K \mid \|\boldsymbol{\beta}_{\mathcal{G}_k}\|_{\text{coop}} > 0\}$, coop-norm triangular inequalities lead to:

$$\begin{aligned}\|\boldsymbol{\beta}\|_{\text{coop}} - \|\hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} + \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} &= \|\boldsymbol{\beta}_{\mathcal{S}(\boldsymbol{\beta})}\|_{\text{coop}} - \|\hat{\boldsymbol{\beta}}^{\text{coop}}_{\mathcal{S}(\boldsymbol{\beta})}\|_{\text{coop}} + \|\boldsymbol{\beta}_{\mathcal{S}(\boldsymbol{\beta})} - \hat{\boldsymbol{\beta}}^{\text{coop}}_{\mathcal{S}(\boldsymbol{\beta})}\|_{\text{coop}} \\ &\leq 2\|\boldsymbol{\beta}_{\mathcal{S}(\boldsymbol{\beta})} - \hat{\boldsymbol{\beta}}^{\text{coop}}_{\mathcal{S}(\boldsymbol{\beta})}\|_{\text{coop}}.\end{aligned}$$

All in all, we obtain that for every $\boldsymbol{\beta} \in \mathbb{R}^p$:

$$\begin{aligned}&\|\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n^2 + \lambda_n \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} \\ &\quad \leq \|\mathbf{X}(\boldsymbol{\beta}^\star - \boldsymbol{\beta})\|_n^2 + 4\lambda_n \|\boldsymbol{\beta}_{\mathcal{S}(\boldsymbol{\beta})} - \hat{\boldsymbol{\beta}}^{\text{coop}}_{\mathcal{S}(\boldsymbol{\beta})}\|_{\text{coop}}.\end{aligned} \tag{A.11}$$

**Second inequality** From optimality conditions given in Theorem 3.4 we deduce that for every group $k \in \{1, \dots, K\}$:

$$\frac{1}{n}\|\mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{coop}})\|_{\text{coop},\infty} \leq \lambda_n. \tag{A.12}$$

Combining inequality (A.12) and the definition of event $\mathcal{A}$ we obtain the following bound:

$$\begin{aligned}\|\mathbf{X}^\intercal \mathbf{X}(\hat{\boldsymbol{\beta}}^{\text{coop}} - \boldsymbol{\beta}^\star)/n\|_{\text{coop},\infty} &\leq \|\mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{coop}})/n\|_{\text{coop},\infty} + \|\mathbf{X}^\intercal \varepsilon / n\|_{\text{coop},\infty} \\ &\leq \lambda_n + \lambda_n/2 \\ &\leq \frac{3}{2}\lambda_n.\end{aligned} \tag{A.13}$$

**Proof of Theorem 3.7**

**Prediction Error** We apply Lemma A.3 with $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ and denote by $\mathcal{S} = \mathcal{S}(\boldsymbol{\beta}^\star)$ the set of true active group indices. Assume $\mathcal{S}$ is of cardinality at

most $s$. On event $\mathcal{A}$, using equivalence relationships between $\|\cdot\|_{\text{coop}}$ and $\|\cdot\|_{\text{coop},2}$, inequality (A.11) becomes:

$$\begin{aligned}\|\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n^2 &\leq 4\lambda_n \|(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})_{\mathcal{S}}\|_{\text{coop}} \\ &\leq 4\lambda_n \sqrt{2s} \|(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})_{\mathcal{S}}\|_{\text{coop},2} \\ &\leq 4\lambda_n \sqrt{2s} \|(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})_{\mathcal{S}}\|.\end{aligned}$$

Now, remark that thanks to (A.11), any coop-Lasso solution satisfies

$$\|(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})_{\mathcal{S}^c}\|_{\text{coop}} \leq 3\|(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})_{\mathcal{S}}\|_{\text{coop}}.$$

Therefore, if Assumption 3.1 is satisfied,

$$\frac{\|\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|}{\sqrt{n}\|(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})_{\mathcal{S}}\|} \geq \kappa(s).$$

In other words,

$$\|(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})_{\mathcal{S}}\| \leq \frac{\|\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|}{\sqrt{n}\kappa(s)},$$

which leads to

$$\begin{aligned}\|X(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n^2 &\leq 4\lambda_n \sqrt{2s} \frac{\|\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n}{\kappa(s)} \\ \|X(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n &\leq 4\lambda_n \frac{\sqrt{2s}}{\kappa(s)} \\ \|X(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n^2 &\leq \frac{32\lambda_n^2 s}{\kappa(s)^2},\end{aligned}$$

**Estimation Error** Similarly, apply the previous lemma with $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$. On event $\mathcal{A}$, inequality (A.11) leads to:

$$\lambda_n \|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} \leq 4\lambda_n \|(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})_{\mathcal{S}}\|_{\text{coop}}$$

Combining Assumption 3.1 and (3.13), we obtain:

$$\begin{aligned}\lambda_n \|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} &\leq 4\lambda_n \sqrt{2s} \|(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})_{\mathcal{S}}\| \\ &\leq 4\lambda_n \sqrt{2s} \frac{\|\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}})\|_n}{\kappa(s)} \\ &\leq \frac{32\lambda_n^2}{\kappa(s)^2} s \\ \|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}^{\text{coop}}\|_{\text{coop}} &\leq \frac{32\lambda_n}{\kappa(s)^2} s.\end{aligned}$$

**Proof of Corollary 3.8: Probability of event $\mathcal{A}$ when groups are all of equal size.**

The choice of $\lambda_n$ is guided by Lemma 5 of S. Negahban and Yu (2012), which we recall here, for $\alpha^* = 2$, for $K$ groups of equal size $p_k = m$

**Lemma A.4** (Lemma 5 of S. Negahban and Yu (2012)) *Suppose that* $\mathbf{X}$ *satisfies the block column normalization condition, and the observation noise* $\varepsilon$ *is sub-Gaussian. Then we have:*

$$\mathbb{P}\left[\max_{k=1,\ldots,K} \|(\mathbf{X}^\intercal\varepsilon)_{\mathcal{G}_k}/n\|_2 \geq 2\frac{\sigma}{\sqrt{n}}(\sqrt{m}+\sqrt{\log K})\right] \leq 2K^{-2}.$$

Recall that event $\mathcal{A}$ is defined by $\{\|\mathbf{X}^\intercal\varepsilon/n\|_{\text{coop},\infty} \leq \lambda_n/2\}$. Now, because for every $z \in \mathbb{R}^p$, $\max(\|z^+\|, \|z^-\|) \leq \|z\|$

$$\begin{aligned}\mathbb{P}\left[\|(\mathbf{X}^\intercal\varepsilon)/n\|_{\text{coop},\infty} \geq \lambda_n/2\right] &= \mathbb{P}\left[\max_{k=1,\ldots,K} \max(\|(\mathbf{X}^\intercal\varepsilon/n)^+_{\mathcal{G}_k}\|, \|(\mathbf{X}^\intercal\varepsilon/n)^-_{\mathcal{G}_k}\|) \geq \lambda_n/2\right] \\ &\leq \mathbb{P}\left[\max_{k=1,\ldots,K} \|(\mathbf{X}^\intercal\varepsilon/n)_{\mathcal{G}_k}\| \geq \lambda_n/2\right]\end{aligned}$$

Choosing $\lambda_n = \frac{\sigma}{\sqrt{n}}(\sqrt{m}+\sqrt{\log K})$, Lemma 5 from S. Negahban and Yu (2012) gives the desired lower bound:

$$\mathbb{P}(\mathcal{A}) \geq 1 - 2K^{-2}$$

## A.2 Proofs for Chapter 4

These proofs have been established by N. Verzelen. We join them to the manuscript for the sake of exhaustivity.

**Additional Notations.** Given a subset $S$, $\Pi_S^{(1)}$ (resp. $\Pi_S^{(2)}$) stands for the orthogonal projection onto the space spanned by the rows of $\mathbf{X}_S^{(1)}$ (resp. $\mathbf{X}_S^{(2)}$). Moreover, $\Pi_{S^\perp}^{(1)}$ denotes the projection along the space spanned by the rows of $\mathbf{X}_{S^c}^{(1)}$.

Besides, we adopt a small change of notations in order to alleviate the equations: $\ell_p$ norms are now denoted by $|\cdot|_p$ instead of $\|\cdot\|_p$, except for the Euclidean norm, which remains denoted by $\|\cdot\|$, with omission of the index 2.

### A.2.1 $F_{S,1}$, $F_{S,2}$ and $F_{S,3}$ distributions (Proposition 4.2)

Let us consider the regression of $Y^{(1)}$ (resp. $Y^{(2)}$) with respect to $X_S^{(1)}$ (resp. $X_S^{(2)}$):

$$\begin{aligned}Y^{(1)} &= X_S^{(1)}\beta_S^{(1)} + \epsilon_S^{(1)} \\ Y^{(2)} &= X_S^{(2)}\beta_S^{(2)} + \epsilon_S^{(2)},\end{aligned}$$

where $X_S^{(1)}\beta_S^{(1)} = \mathbb{E}[Y|X_S^{(1)}]$ and $X_S^{(2)}\beta_S^{(2)} = \mathbb{E}[Y|X_S^{(2)}]$ a.s. We note $(\sigma_S^{(1)})^2 = \text{Var}(\epsilon_S^{(1)}) = \text{Var}[Y^{(1)}|X_S^{(1)}]$ and $(\sigma_S^{(2)})^2 = \text{Var}(\epsilon_S^{(2)}) = \text{Var}[Y^{(2)}|X_S^{(2)}]$. Under $\mathcal{H}_{0,S}$, we have $\beta_S^{(1)} = \beta_S^{(2)}$ and $\sigma_S^{(1)} = \sigma_S^{(2)}$. For the sake of simplicity, we write $\beta_S$, $\sigma_S$ for these two quantities.

Define the random variable $T_1$ and $T_2$ as

$$T_1 = \frac{\|\Pi^{(1)}_{S^\perp}\boldsymbol{\epsilon}^{(1)}_S\|^2}{(n_1 - |S|)\left(\sigma^{(1)}_S\right)^2}, \qquad T_2 = \frac{\|\Pi^{(2)}_{S^\perp}\boldsymbol{\epsilon}^{(2)}_S\|^2}{(n_2 - |S|)\left(\sigma^{(2)}_S\right)^2}.$$

Conditionnally to $\mathbf{X}$, $T_1/T_2$ follows a Fisher distribution with $(n_1 - |S|, n_2 - |S|)$ degrees of freedom. Observing that

$$F_{S,1} = -2 + \frac{T_1}{T_2}\frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)} + \frac{T_2}{T_1}\frac{n_1(n_2 - |S|)}{n_2(n_1 - |S|)}$$

allows us to prove the first assertion of Proposition 4.2.

Let us turn to the second statistic $F_{2,S}$

$$F_{S,2} = \frac{n_1}{n_2(n_1 - |S|)}\frac{U}{T_1},$$

where

$$U = \frac{\|\mathbf{X}^{(2)}_S(\mathbf{X}^{(2)\intercal}_S\mathbf{X}^{(2)}_S)^{-1}\mathbf{X}^{(2)\intercal}_S\boldsymbol{\epsilon}^{(2)}_S - \mathbf{X}^{(2)}_S(\mathbf{X}^{(1)\intercal}_S\mathbf{X}^{(1)}_S)^{-1}\mathbf{X}^{(1)\intercal}_S\boldsymbol{\epsilon}^{(1)}_S\|^2}{\sigma^2_S}$$

Conditionnally to $\mathbf{X}$, $U$ is independent of $T_1$ since $T_1$ is a function of $\Pi^{(1)}_{S^\perp}\boldsymbol{\epsilon}^{(1)}_S$ while $U$ is a function of $(\boldsymbol{\epsilon}^{(2)}_S, \Pi^{(1)}_S\boldsymbol{\epsilon}^{(1)}_S)$. Furthermore, $U$ is the squared norm of a centered Gaussian vector with covariance

$$\mathbf{X}^{(2)}_S\left[(\mathbf{X}^{(1)\intercal}_S\mathbf{X}^{(1)}_S)^{-1} + (\mathbf{X}^{(2)\intercal}_S\mathbf{X}^{(2)}_S)^{-1}\right]\mathbf{X}^{(2)\intercal}_S.$$

### A.2.2 Power of $T^B_{\mathcal{S}}$ for a Deterministic Collection $\mathcal{S}$ (Theorem 4.7)

The objective is to exhibit a subset over which the power of $T^B_{\mathcal{S}}$ is larger than $1 - \delta$. This subset is such that the distance between the two sample-specific distributions is large enough that we can actually reject the null hypothesis with large probability. As exposed in Section 4.2, the distance that naturally arises is the sum $\mathcal{K}_1 + \mathcal{K}_2$, which forms a semidistance between $(\beta^{(1)}, \sigma_1)$ and $(\beta^{(2)}, \sigma_2)$. We recall that $\mathcal{K}_1 + \mathcal{K}_2$ decomposes into three terms which correspond respectively to the statistics $F_{S,1}$, $F_{S,2}$ and $F_{S,3}$:

$$\begin{aligned} 2(\mathcal{K}_1(S) + \mathcal{K}_2(S)) = &\left(\frac{\sigma^{(1)}_S}{\sigma^{(2)}_S}\right)^2 + \left(\frac{\sigma^{(2)}_S}{\sigma^{(1)}_S}\right)^2 - 2 \\ &+ \frac{\|\beta^{(2)}_S - \beta^{(1)}_S\|^2_{\Sigma^{(2)}}}{(\sigma^{(2)}_S)^2} \\ &+ \frac{\|\beta^{(2)}_S - \beta^{(1)}_S\|^2_{\Sigma^{(1)}}}{(\sigma^{(1)}_S)^2}. \end{aligned} \tag{A.14}$$

The proof is split into five main lemmas. First, we upper bound $\widetilde{Q}^{-1}_{1,|S|}(x|\mathbf{X}_S)$, $\widetilde{Q}^{-1}_{2,|S|}(x|\mathbf{X}_S)$, and $\widetilde{Q}^{-1}_{3,|S|}(x|\mathbf{X}_S)$ in Lemmas A.5,A.6 and A.7.

Then, we control the deviations of $F_{S,1}$, $F_{S,2}$, and $F_{S,3}$ around the three corresponding terms of (A.14) under $\mathcal{H}_{1,S}$ in Lemmas A.8 and A.9

Based on these lemmas, we can provide conditions on each of the three terms of $\mathcal{K}_1 + \mathcal{K}_2$ in order for the power of $T_S^B$ to exceed $1-\delta$.

Throughout this proof, we assume that , and $S \leq (n_1 \wedge n_2)/2$

$$\log(12/\delta) < 2^{-13}(n_1 \wedge n_2), \quad \log(1/\alpha_S) \leq 2^{-10}(n_1 \wedge n_2)\,, \tag{A.15}$$

for any $S \in \mathbf{S}'$. These two conditions allow to fix the constant $L$ in the statement (4.16) of Theorem 4.7.

**Lemma A.5** (Upper-bound on $\widetilde{Q}^{-1}_{1,|S|}(x|\mathbf{X}_S)$ ) *Consider some $0 < x < 1$ such that $8\log(2/x) \leq n_1 \wedge n_2$ . For any subset $S$ of size smaller than $(n_1 \wedge n_2)/2$, we have*

$$\widetilde{Q}^{-1}_{1,|S|}(x|\mathbf{X}_S) \leq 2^{12}\left\{\left(\frac{|S|(n_1-n_2)}{n_1 n_2}\right)^2 + \log(2/x)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right\}\,. \tag{A.16}$$

We recall that $a = (a_1, \ldots, a_{|S|})$ denotes the positive eigenvalues of

$$\frac{n_1}{n_2(n_1-|S|)}\mathbf{X}_S^{(2)}\left[(\mathbf{X}_S^{(1)\intercal}\mathbf{X}_S^{(1)})^{-1} + (\mathbf{X}_S^{(2)\intercal}\mathbf{X}_S^{(2)})^{-1}\right]\mathbf{X}_S^{(2)\intercal}\,.$$

**Lemma A.6** (Upper-bound on $\widetilde{Q}^{-1}_{2,|S|}(x|\mathbf{X}_S)$ ) *If $|a|_1 < u \leq (n_1-|S|)|a|_\infty$ and if $|S| \leq 2^{-6}n_1$,*

$$\log\left[\widetilde{Q}_{2,|S|}(u|\mathbf{X}_S)\right] \leq -\frac{(u-|a|_1)^2}{4\left[|a|_\infty(u-|a|_1)+|a|_2^2\right]} + \frac{(u-|a|_1)u^3}{2(n_1-|S|)\left[|a|_\infty(u-|a|_1)+|a|_2^2\right]^2}\,.$$

*For any $0 < x < 1$, satisfying*

$$2^9\log(1/x) \leq n_1 - |S|\,, \tag{A.17}$$

*we have the following upper bound*

$$\widetilde{Q}^{-1}_{2,|S|}(x|\mathbf{X}_S) \leq |a|_\infty\left[|S| + 2\sqrt{2|S|\log(1/x)} + 8\log(1/x)\right]\,. \tag{A.18}$$

**Lemma A.7** (Upper-bound on $|a|_\infty$) *Consider $\delta$ a positive number sastifying $\log(4/\delta) < (n_1 \wedge n_2)2^{-7}$. With probability larger than $1-\delta/2$, we have*

$$|a|_\infty \;\leq\; 100\left[\frac{1}{n_2} + \frac{\varphi_{\max}\left\{\sqrt{\Sigma_S^{(2)}}(\Sigma_S^{(1)})^{-1}\sqrt{\Sigma_S^{(2)}}\right\}}{n_1}\right]\,.$$

**Lemma A.8** (Deviations of $F_{S,1}$) *Assume that $\log(1/\delta) \leq 2^{-10}(n_1 \wedge n_2)$. With probability larger than $1-\delta$, we have*

$$F_{S,1} \geq 2^{-4}\left(\frac{(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2}{\sigma_S^{(1)}\sigma_S^{(2)}}\right)^2 - 2^{14}\left[|S|^2\left(\frac{1}{n_1^2}+\frac{1}{n_2^2}\right) + \log(\frac{1}{\delta})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right] \tag{A.19}$$

**Lemma A.9** (Deviations of $F_{S,2}$) *Assume that*

$$\log(12/\delta) < 2^{-11}(n_1 \wedge n_2)\,. \tag{A.20}$$

*With probability larger than* $1-\delta/2$, *we have*

$$F_{S,2} \geq \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|^2_{\Sigma^{(2)}}}{8(\sigma_S^{(1)})^2} - \frac{\log(6/\delta)}{n_2} 200 \left[\frac{\sigma^2_{(2),|S|}}{\sigma^2_{(1),|S|}} + \varphi_S\right]\,, \tag{A.21}$$

*where* $\|.\|_{\Sigma^{(2)}}$ *is the euclidean norm relative to* $\Sigma^{(2)}$.

Consider some $S \in \mathbf{S}'$. Combining Lemmas A.5 and A.8, $\widetilde{Q}_{1,|S|}(F_{S,1}|\mathbf{X}_S) \leq \alpha_S$ holds with probability larger than $1-\delta$ if

$$\frac{((\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2)^2}{(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2} \geq 2^{19}\left[|S|^2\left(\frac{1}{n_1^2}+\frac{1}{n_2^2}\right) + \log[1/(\alpha_S\delta)]\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right]\,.$$

Similarly, combining Lemmas A.6, A.7 and A.9, $\widetilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S) \leq \alpha_S$ with probability larger than $1-\delta$ if

$$\begin{aligned}\frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|^2_{\Sigma^{(2)}}}{(\sigma_S^{(1)})^2} \geq & 1600\,(\varphi_S+1)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)[|S| + 5\log\{6/(\delta\alpha_S)\}] \\ & + \frac{1600}{n_2}\left(\frac{\sigma^{(2)}_{|S|}}{\sigma^{(1)}_{|S|}}\right)^2 \log\left(\frac{6}{\delta}\right)\,.\end{aligned}$$

A symmetric result holds for $\widetilde{Q}_{3,|S|}(F_{S,3}|\mathbf{X}_S)$.

Consequently, $\widetilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S) \wedge \widetilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S) \wedge \widetilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S) \leq \alpha_S$ with probability larger than $1-\delta$ if

$$\begin{aligned}\mathcal{K}_1(S) + \mathcal{K}_2(S) \geq & \; 2^{21}\varphi_S\left(\frac{1}{n_1}+\frac{1}{n_2}\right)[|S| + 6\log\{6/(\alpha_S\delta)\}] \\ & + 1600\log(6/\delta)\left[\frac{1}{n_1}+\frac{1}{n_2}\right]\left[\left(\frac{\sigma^{(2)}_{|S|}}{\sigma^{(1)}_{|S|}}\right)^2 + \left(\frac{\sigma^{(1)}_{|S|}}{\sigma^{(2)}_{|S|}}\right)^2\right]\,.\end{aligned}$$

Since $6400\log(6/\delta) \leq n_1 \wedge n_2$, the last condition is fulfilled if

$$\mathcal{K}_1(S) + \mathcal{K}_2(S) \geq 2^{22}\varphi_S\left(\frac{1}{n_1}+\frac{1}{n_2}\right)[|S| + 7\log\{6/(\alpha_S\delta)\}]\,. \tag{A.22}$$

We now proceed to the proof of the five previous lemmas.

*Proof of Lemma A.5.* Let $u \in (0,1)$ and $\bar{F}^{-1}_{D,N}(u)$ be the $1-u$ quantile of a Fisher random variable with $D$ and $N$ degrees of freedom. According to Baraud et al. (2003), we have

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

Let us assume that $4/N \log(1/u) \leq 1$. By convexity of the exponential function it holds that

$$\bar{F}^{-1}_{D,N}(u) \;\leq\; 1 + 2\sqrt{\left(\frac{1}{D}+\frac{1}{N}\right)\log\left(\frac{1}{u}\right)}\left(\frac{4}{D}+\frac{8}{N}\right)\log\left(\frac{1}{u}\right) .$$

Under the hypothesis $\mathcal{H}_0$,

$$\frac{T_1}{T_2}\,\frac{n_1(n_2-|S|)}{n_2(n_1-|S|)} \sim \text{Fisher}(n_1-|S|, n_2-|S|) .$$

Consider some $x > 0$ such that $[4/(n_1-|S|) \vee 4/(n_2-|S|)]\log(2/x) \leq 1$. Then, with probability larger than $1-x/2$ we have,

$$\begin{aligned}\frac{T_1}{T_2} &\leq \left(1+\frac{|S|(n_1-n_2)}{n_1(n_2-|S|)}\right)\left(1+8\sqrt{\frac{\log(2/x)}{n_1-|S|}}+8\sqrt{\frac{\log(2/x)}{n_2-|S|}}\right) \leq 18 \\ &\leq \left(1+\frac{|S|(n_1-n_2)}{n_1(n_2-|S|)}\right)\left(1+12\sqrt{\frac{\log(2/x)}{n_1}}+12\sqrt{\frac{\log(2/x)}{n_2}}\right) ,\end{aligned}$$

since $|S| \leq (n_1 \wedge n_2)/2$. Similarly, with probability at least $1-x/2$, we have

$$\frac{T_2}{T_1} \leq \left[\left(1+\frac{|S|(n_2-n_1)}{n_2(n_1-|S|)}\right)\left(1+12\sqrt{\frac{\log(2/x)}{n_1}}+12\sqrt{\frac{\log(2/x)}{n_2}}\right)\right] \wedge 18 \tag{A.23}$$

Depending on the sign of $T_1/T_2 - 1$, we apply one the two following identities:

$$\frac{T_1}{T_2}+\frac{T_2}{T_1}-2 = \left(\frac{T_1}{T_2}-1\right)^2\frac{T_2}{T_1} , \qquad \frac{T_1}{T_2}+\frac{T_2}{T_1}-2 = \left(\frac{T_2}{T_1}-1\right)^2\frac{T_1}{T_2} .$$

Combining the different bounds, we conclude that with probability larger than $1-x$,

$$\frac{T_1}{T_2}+\frac{T_2}{T_1}-2 \;\leq\; 2^{12}\left[\left(\frac{|S|(n_1-n_2)}{n_1 n_2}\right)^2+\log(2/x)\frac{n_1+n_2}{n_1 n_2}\right] .$$

□

*Proof of Lemma A.6.* As in the proof of Proposition 4.3, we note $N = n_1 - |S|$. Recall that $\widetilde{Q}_{2,|S|}(x|\mathbf{X}_S)$ is defined as $\inf_{0<\lambda<|a|_\infty/2}\exp\psi_x(\lambda) = \exp\psi_x(\lambda^\star)$. We start by upper-bounding $\psi_u(\lambda^\star)$, which in other words proves the first upper-bound on the logarithm of the tail probability $\log\widetilde{Q}_{2,|S|}(u|\mathbf{X}_S)$. We then exhibit a value $u_x$ such that $\psi_{u_x}(\lambda^\star) \leq \log x$.

**Upper-bound on the tail probability.** Since the equation (4.22) is increasing with respect to $\lambda$ and with respect to $N$, $\lambda^*$ decreases with $N$. Consequently,

$$\lambda^* \leq \lambda_+ := \frac{u-|a|_1}{2\left[|a|_\infty(u-|a|_1)+|a|_2^2\right]} .$$

By convexity, $1-\sqrt{1-x} \geq x/2$ for any $0 \leq x \leq 1$. Applying this inequality, we upper bound bound $\sqrt{\Delta}$ and derive that

$$\lambda^* \geq \lambda_- := \frac{u - |a|_1}{2\left[|a|_\infty(u-|a|_1) + |a|_2^2 + \frac{|a|_1 u}{N}\right]} \cdot$$

Since $u \leq N|a|_\infty$, $2\lambda^* u \leq N$. Observing that $-\log(1-2x)/2 \leq x + x^2/(1-2x)$ for any $0 < x < 1/2$, we derive

$$\begin{aligned}
\psi_u(\lambda^*) &\leq |a|_1\lambda_+ + \frac{\lambda_+^2|a|_2^2}{1-2|a|_\infty\lambda_+} - \lambda^* u + 2\frac{(\lambda^*)^2u^2}{N} \\
&\leq -\frac{(u-|a|_1)^2}{4\left[|a|_\infty(u-|a|_1)+|a|_2^2\right]} + \frac{2\lambda_+^2u^2}{N} + (\lambda_+ - \lambda_-)u \\
&\leq -\frac{(u-|a|_1)^2}{4\left[|a|_\infty(u-|a|_1)+|a|_2^2\right]} + \frac{(u-|a|_1)u^3}{2N\left[|a|_\infty(u-|a|_1)+|a|_2^2\right]^2} \cdot
\end{aligned}$$

**Upper-bound on the quantile.** Let us turn to the upper bound of $\widetilde{Q}^{-1}_{2,|S|}(x|\mathbf{X}_S)$. Consider $u_x$ the solution larger than $|a|_1$ of the equation

$$\frac{(u-|a|_1)^2}{4\left[|a|_\infty(u-|a|_1)+|a|_2^2\right]} = 2\log(1/x)\,,$$

and observe that

$$2|a|_2\sqrt{\log(1/x)} \leq u_x - |a|_1 \leq 2\sqrt{2}|a|_2\sqrt{\log(1/x)} + 8|a|_\infty\log(1/x)\,.$$

By Condition (A.17), $u_x \leq N|a|_\infty$. We now prove that $\psi_{u_x \vee 2|a|_1}(\lambda^*) \leq \log x$.

If $u_x \geq 2|a|_1$, then $u_x^3 \leq 8(u_x - |a|_1)^3$ and it follows that

$$\psi_{u_x}(\lambda^*) \leq \log(1/x)\left[-2 + \frac{2^8\log(1/x)}{N}\right] \leq -\log(1/x)$$

by Condition (A.17).

If $u_x \leq 2|a|_1$, then $|a|_1^2/(|a|_\infty|a|_1 + |a|_2^2) \geq 8\log(1/x)$ and

$$\psi_{u_x\vee|a|_1}(\lambda^*) \leq -\frac{|a|_1^2}{4\left[|a|_\infty|a|_1+|a|_2^2\right]}\left[1 - \frac{2^4|a|_1^2}{N\left[|a|_\infty|a|_1+|a|_2^2\right]}\right] \leq -\log(1/x)\,,$$

since $|S| \leq 2^{-6}n_1$.

All in all, we conclude that

$$\widetilde{Q}^{-1}_{2,|S|}(x|\mathbf{X}_S) \leq u_x \vee 2|a|_1 \leq |a|_1 + \left[2\sqrt{2}|a|_2\sqrt{\log(1/x)} + 8|a|_\infty\log(1/x)\right] \vee |a|_1\,.$$

□

*Lemma A.7.* Upon defining $\mathbf{Z}_S^{(1)} = \sqrt{\Sigma_S^{(1)}}^{-1}\mathbf{X}_S^{(1)}$ and $\mathbf{Z}_S^{(2)} = \sqrt{\Sigma_S^{(2)}}^{-1}\mathbf{X}_S^{(2)}$, it follows that $\mathbf{Z}_S^{(1)}$ and $\mathbf{Z}_S^{(2)}$ follow standard Gaussian distributions.

$$
\begin{aligned}
|a|_\infty &\leq \frac{n_1}{n_2(n_1-|S|)}\left[1+\varphi_{\max}\left\{\mathbf{Z}_S^{(2)\intercal}\sqrt{\Sigma_S^{(2)}(\Sigma_S^{(1)})^{-1}}\left(\mathbf{Z}_S^{(1)\intercal}\mathbf{Z}_S^{(1)}\right)^{-1}\sqrt{(\Sigma_S^{(1)})^{-1}\Sigma_S^{(2)}}\mathbf{Z}_S^{(2)}\right\}\right] \\
&\leq \frac{2}{n_2}+2\frac{\varphi_{\max}[\mathbf{Z}_S^{(2)\intercal}\mathbf{Z}_S^{(2)}]}{n_2\varphi_{\max}[\mathbf{Z}_S^{(1)\intercal}\mathbf{Z}_S^{(1)}]}\varphi_{\max}\left[\sqrt{\Sigma_S^{(2)}}(\Sigma_S^{(1)})^{-1}\sqrt{\Sigma_S^{(2)}}\right] .
\end{aligned}
$$

In order to conclude, we control the largest and the smallest eigenvalues of Standard Wishart matrices by applying Lemma A.16. □

*Lemma A.8.* By symmetry, we may assume that $(\sigma_S^{(1)})^2/(\sigma_S^{(2)})^2 \geq 1$.

**CASE 1**. Suppose that $T_1/T_2 \geq 1$.

$$
\begin{aligned}
-2+\frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2}\frac{T_1}{T_2}+\frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2}\frac{T_2}{T_1} &\geq \frac{((\sigma_S^{(1)})^2-(\sigma_S^{(2)})^2)^2}{(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2}+\frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2}\left(\frac{T_1}{T_2}-1\right)+\frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2}\left(\frac{T_2}{T_1}-1\right) \\
&\geq \frac{((\sigma_S^{(1)})^2-(\sigma_S^{(2)})^2)^2}{(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2} . \qquad \text{(A.24)}
\end{aligned}
$$

**CASE 2**. Suppose that $T_1/T_2 \leq 1$.

$$
\begin{aligned}
-2+\frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2}\frac{T_1}{T_2}+\frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2}\frac{T_2}{T_1} &= \left(\frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2}-\frac{T_2}{T_1}\right)^2\frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2}\frac{T_1}{T_2} \\
&\geq \frac{T_1}{T_2}\frac{((\sigma_S^{(1)})^2-(\sigma_S^{(2)})^2)^2}{4(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2}\mathbf{1}_{\frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2}-1\geq 2\left(\frac{T_2}{T_1}-1\right)} .
\end{aligned}
$$

We now need to control the deviations of $T_2/T_1$. Using the bound (A.23), we get

$$
\frac{T_2}{T_1}\leq\left(1+\frac{|S|(n_2-n_1)}{n_2(n_1-|S|)}\right)\left(1+12\sqrt{\frac{\log(1/\delta)}{n_1}}+12\sqrt{\frac{\log(1/\delta)}{n_2}}\right) ,
$$

with probability larger than $1-\delta$. Since $|S|\leq(n_1\wedge n_2)/2$, we derive that

$$
\frac{T_2}{T_1}-1\leq\frac{2|S|}{n_1}+24\sqrt{\frac{\log(1/\delta)}{n_1}}+24\sqrt{\frac{\log(1/\delta)}{n_2}}\leq 3 .
$$

In conclusion, we have

$$
-2+\frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2}\frac{T_1}{T_2}+\frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2}\frac{T_2}{T_1}\geq\frac{((\sigma_S^{(1)})^2-(\sigma_S^{(2)})^2)^2}{16(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2} , \qquad \text{(A.25)}
$$

with probability larger than $1-\delta$, as long as

$$
\frac{((\sigma_S^{(1)})^2-(\sigma_S^{(2)})^2)^2}{(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2}\geq 2^{14}\left[\frac{|S|^2}{n_1^2}+\frac{|S|^2}{n_2^2}+\log(1/\delta)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right] . \qquad \text{(A.26)}
$$

Combining (A.24), (A.25), and (A.26), we derive

$$-2+\frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2}\frac{T_1}{T_2}+\frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2}\frac{T_2}{T_1}\geq\frac{((\sigma_S^{(1)})^2-(\sigma_S^{(2)})^2)^2}{16(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2}-2^{14}\left[\frac{|S|^2}{n_1^2}+\log(1/\delta)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right],$$

with probability larger than $1-\delta$. □

*Lemma A.9.* We want to lower bound the random variable $F_{S,2} = \frac{Rn_1}{(\sigma_S^{(1)})^2T_1(n_1-|S|)}$ where $R$ is defined

$$R := \|\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})+\Pi_S^{(2)}\boldsymbol{\epsilon}_S^{(2)}-(\sigma_S^{(1)})^2U_1\|^2/n_2\,.$$

Let us first work conditionally to $\mathbf{X}_S^{(1)}$ and $\mathbf{X}_S^{(2)}$. Upon defining the Gaussian vector $W$ by

$$W\sim\mathcal{N}\left[0,(\sigma_S^{(2)})^2\Pi_S^{(2)}+(\sigma_S^{(1)})^2\mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\intercal}\mathbf{X}_S^{(1)})^{(-1)}\mathbf{X}_S^{(2)\intercal}\right],$$

we get $R = \|\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})+W\|^2/n_2$. We have the following lower bound:

$$\begin{aligned}R &\geq \left(\|\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})\|+\left\langle W,\frac{\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})}{\|\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})\|}\right\rangle\right)^2/n_2\\ &\geq \frac{\|\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})\|^2}{2n_2}-\frac{1}{n_2}\left\langle W,\frac{\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})}{\|\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})\|}\right\rangle^2\end{aligned}$$

The random variable $\|\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})\|^2/\|\beta_S^{(2)}-\beta_S^{(1)}\|^2_{\Sigma^{(2)}}$ follows a $\chi^2$ distribution with $n_2$ degrees of freedom. Conditionally to $\mathbf{X}_S$, $\left\langle W,\frac{\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})}{\|\mathbf{X}_S^{(2)}(\beta_S^{(2)}-\beta_S^{(1)})\|}\right\rangle^2$ is proportional to $\chi^2$ distributed random variable with 1 degree of freedom and its variance is smaller than $(\sigma_S^{(2)})^2+\varphi_{\max}(\mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\intercal}\mathbf{X}_S^{(1)})^{(-1)}\mathbf{X}_S^{(2)\intercal})(\sigma_S^{(1)})^2$.

Applying Lemma A.15, we derive that with probability larger than $1-x/6$,

$$\begin{aligned}R &\geq \frac{\|\beta_S^{(2)}-\beta_S^{(1)}\|^2_{\Sigma^{(2)}}}{2}\left[1-2\sqrt{\frac{\log(12/x)}{n_2}}\right]\\ &- 4\frac{\log(12/x)}{n_2}\left[(\sigma_S^{(2)})^2+(\sigma_S^{(1)})^2\varphi_{\max}(\mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\intercal}\mathbf{X}_S^{(1)})^{(-1)}\mathbf{X}_S^{(2)\intercal})\right].\end{aligned}$$

Using the upper bound $|S|\leq(n_1\wedge n_2)/2$ and Lemma A.16, we control the last term

$$\varphi_{\max}\left[\mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\intercal}\mathbf{X}_S^{(1)})^{(-1)}\mathbf{X}_S^{(2)\intercal}\right]\leq 50\varphi_S\,,$$

with probability larger than $1-2\exp[-(n_1\wedge n_2)0.04^2/2]$. By condition (A.20),

$$R \geq \frac{\|\beta_S^{(2)}-\beta_S^{(1)}\|^2_{\Sigma^{(2)}}}{4}-\frac{\log(12/\delta)}{n_2}200\left[\sigma^2_{(2),|S|}+\sigma^2_{(1),|S|}\varphi_S\right], \quad \text{(A.27)}$$

with probability larger than $1-\delta/2$.

Let us now upper bound the random variable $T_1(n_1-|S|)/n_1$. Since $(n_1-S)T_1$ follows a $\chi^2$ distribution with $n_1-|S|$ degrees of freedom, we derive from Lemma A.15 that

$$T_1(n_1-|S|)/n_1 \;\leq\; 1+2\sqrt{\frac{\log(6/\delta)}{n_1}}+\frac{2}{n_1}\log(6/\delta)\leq 2\,, \quad \text{(A.28)}$$

with probability larger than $1-\delta/6$. Gathering (A.27) and (A.28), we conclude that

$$F_{S,2}\geq \frac{\|\beta_S^{(2)}-\beta_S^{(1)}\|^2_{\Sigma^{(2)}}}{8(\sigma_S^{(1)})^2}-\frac{\log(6/\delta)}{n_2}2^{16}\left[\frac{\sigma^2_{(2),|S|}}{\sigma^2_{(1),|S|}}+\varphi_S\right]\,,$$

with probability larger than $1-\delta/2$.

□

### A.2.3 Power of $T^B_{\mathcal{S}_{\leq k}}$ (Proposition 4.6)

This proposition is a straightforward corollary of Theorem 4.7. If the constant $L_1$ in (4.12) is large enough, then condition (4.16) is fulfilled for any subsets $S$ of size less than $k\wedge k_*$. Assume first that $S^{\cup(1,2)}\neq\emptyset$ or $S^{\cap(1,2)}\neq\emptyset$. Applying Theorem 4.7, we derive that $T^B_\alpha$ rejects $\mathcal{H}_0$ with probability larger than $1-\delta$ when $\mathcal{K}_1(S^{\cup(1,2)})+\mathcal{K}_2(S^{\cup(1,2)})\geq\Delta(S^{\cup(1,2)})$ or $\mathcal{K}_1(S^{\cap(1,2)})+\mathcal{K}_2(S^{\cap(1,2)})\geq\Delta(S^{\cap(1,2)})$. Working out $\mathcal{K}_1(S^{\cup(1,2)})$ and $\mathcal{K}_1(S^{\cap(1,2)})$ allows us to conclude. If $S^{\cup(1,2)}=\emptyset$ or $S^{\cap(1,2)}=\emptyset$, then we consider any subset of size 1.

### A.2.4 Power of $T^B_{\widehat{\mathcal{S}}_{\text{Lasso}}}$ (Theorem 4.8)

For simplicity, we assume in the sequel that $\beta^{(1)}\neq 0$ or $\beta^{(2)}\neq 0$, the case $\beta^{(1)}=\beta^{(2)}=0$ being handled by any set $S\in\widehat{\mathbf{S}}_{\text{Lasso}}$ of size 1.

Given a matrix $\mathbf{X}$, an integer $k$, and a number $M$, the largest and smallest eigenvalues of order $k$ and the compatibility constant $\kappa[M,k,\mathbf{X}]$ (see Raskutti et al. (2010)) are respectively defined by

$$\begin{aligned}\Phi_{k,+}(\mathbf{X}) &= \sup_{\theta,1\leq|\theta|_0\leq k}\frac{\|\mathbf{X}\theta\|^2}{\|\theta\|^2}\,,\quad \Phi_{k,-}(\mathbf{X})=\inf_{\theta,1\leq|\theta|_0\leq k}\frac{\|\mathbf{X}\theta\|^2}{\|\theta\|^2}\,,\\ \kappa[M,k,\mathbf{X}] &= \min_{T,\theta:\ |T|\leq k,\ \theta\in\mathcal{C}(M,T)}\left\{\frac{\|\mathbf{X}\theta\|}{\|\theta\|}\right\}\,,\end{aligned}$$

where $\mathcal{C}(M,T)=\{\theta:\ |\theta_{T^c}|_1<M|\theta_T|_1\}$. Define $k_*$ as the largest integer that satisfies

$$2(k_*+1)\log(p)\leq 2^{-15}(n_1\wedge n_2)\,. \quad \text{(A.29)}$$

We also consider the quantity

$$\gamma_{\Sigma^{(1)},\Sigma^{(2)},\beta}:=\frac{\bigvee_{i=1,2}\Phi^2_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2}\Phi_{k_*,-}(\sqrt{\Sigma^{(i)}})\bigwedge_{i=1,2}\kappa^2[10,|\beta^{(1)}|_0+|\beta^{(2)}|_0,\sqrt{\Sigma^{(i)}}]}\,,$$

that measures the closeness to orthogonality of $\Sigma^{(1)}$ and $\Sigma^{(2)}$. The next proposition is a sharper result than Theorem 4.8.

**Proposition A.10** *The following positive numerical constants $c_1$, $c_2$, and $c_3$ are introduced in the proof of Lemma A.11 below Assume that*

$$\log\left[(24 \vee c_1)/(\alpha\delta)\right] < (2^{-14} \wedge c_2)(n_1 \wedge n_2)\,. \tag{A.30}$$

*The hypothesis $\mathcal{H}_0$ is rejected by $T_\alpha^B$ with probability larger than $1-\delta$ for any $(\beta^{(1)}, \beta^{(2)})$ satisfying*

$$\frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2\left[10, 2|S^{\cup(1,2)}|, \sqrt{\Sigma^{(i)}}\right]} |S^{\cup(1,2)}| \le \left(2^{-14} \wedge c_3^{-1}\right) k_*\,. \tag{A.31}$$

*and*

$$\mathcal{K}_1 + \mathcal{K}_2 \ge L\gamma_{\Sigma^{(1)},\Sigma^{(2)},\beta} \frac{\left(|S^{\cup(1,2)}| \vee 1\right)\log(p) + \log\{1/(\alpha\delta)\}}{n_1 \wedge n_2}\,.$$

This proof of Proposition A.10 is divided in two main steps. First, we prove that with large probability the collection $\widehat{\mathbf{S}}_{\text{Lasso}}$ contains some set $\widehat{S}_\lambda$ close to

$$S^{\cup(1,2)} = \text{supp}(\beta^{(1)}) \cup \text{supp}(\beta^{(2)})\,.$$

Then, the statistics $(F_{\widehat{S}_\lambda,1}, F_{\widehat{S}_\lambda,2}, F_{\widehat{S}_\lambda,3})$ allow to reject $\mathcal{H}_0$ with large probability. Recall that the collection $\widehat{\mathbf{S}}_{\text{Lasso}}$ is based on the Lasso regularization path of the following heteroscedastic Gaussian linear model,

$$\begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)}/\sqrt{n_1} & \mathbf{X}^{(1)}/\sqrt{n_1} \\ \mathbf{X}^{(2)}/\sqrt{n_2} & -\mathbf{X}^{(2)}/\sqrt{n_2} \end{bmatrix} \begin{bmatrix} \theta^{(1)} \\ \theta^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \end{bmatrix} \tag{A.32}$$

which we denote for short:

$$\mathbf{Y} = \mathbf{W}\theta_0 + \boldsymbol{\epsilon}.$$

Given a tuning parameter $\lambda$, $\widehat{\theta}_\lambda$ refers to the Lasso estimator of $\theta$:

$$\widehat{\theta}_\lambda = \arg\inf_{\theta\in\mathbb{R}^{2p}} \|\mathbf{Y} - \mathbf{W}\theta\|^2 + \lambda|\theta|_1\,.$$

In order to analyze the Lasso solution $\widehat{\theta}_\lambda$, we need to control how $\mathbf{W}$ acts on sparse vectors.

**Lemma A.11** (Control of the design $\mathbf{W}$) *The event*

$$\begin{aligned}\mathcal{A} \;:=\; & \left\{\forall\theta \text{ s.t. } |\theta|_0 \le k_*,\ 1/2 \le \frac{\|\mathbf{X}^{(1)}\theta\|^2}{n_1\|\theta\|^2_{\Sigma^{(1)}}} \le 2 \text{ and } 1/2 \le \frac{\|\mathbf{X}^{(2)}\theta\|^2}{n_2\|\theta\|^2_{\Sigma^{(2)}}} \le 2\right\}\,. \\ & \bigcap \left\{\frac{\kappa\left[10, |\beta^{(1)}|_0 + |\beta^{(2)}|_0, \mathbf{X}^{(1)}/\sqrt{n_1}\right]}{\kappa\left[10, |\beta^{(1)}|_0 + |\beta^{(2)}|_0, \sqrt{\Sigma^{(1)}}\right]} \bigwedge \frac{\kappa\left[10, |\beta^{(1)}|_0 + |\beta^{(2)}|_0, \mathbf{X}^{(2)}/\sqrt{n_1}\right]}{\kappa\left[10, |\beta^{(1)}|_0 + |\beta^{(2)}|_0, \sqrt{\Sigma^{(2)}}\right]} \ge 2^{-3}\right\}\end{aligned}$$

*has large probability $\mathbb{P}[\mathcal{A}] \ge 1 - \delta/4$. Furthermore, on the event $\mathcal{A}$,*

$$\begin{aligned}\Phi_{k,+}(\mathbf{W}) &\le 4\left[\Phi_{k,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{k,+}(\sqrt{\Sigma^{(2)}})\right]\,, \\ \Phi_{k,-}(\mathbf{W}) &\ge \Phi_{k,-}(\sqrt{\Sigma^{(1)}}) \wedge \Phi_{k,-}(\sqrt{\Sigma^{(2)}})\,,\end{aligned}$$

*for any $k \le k_*$.*

The following lemma is a slight variation of Theorem 14 in Koltchinski et al. (2011) and Lemma 3.2 in Giraud et al. (2012).

**Lemma A.12** (Behavior of the Lasso estimator $\widehat{\theta}_\lambda$) *The event*

$$\mathcal{B} = \left\{ |\mathbf{W}^T \boldsymbol{\epsilon}|_\infty \leq 2(\sigma_1 \vee \sigma_2)\sqrt{2\Phi_{1,+}(\mathbf{W})\log(p)} \right\}$$

*occurs with probability larger than* $1 - 1/p$. *Assume that*

$$\lambda \geq 8(\sigma_1 \vee \sigma_2)\sqrt{2\Phi_{1,+}(\mathbf{W})\log(p)}\,.$$

*Then, on the event* $\mathcal{A} \cap \mathcal{B}$ *we have*

$$\|\mathbf{W}(\widehat{\theta}_\lambda - \theta_0)\|^2 \leq \frac{2^6\lambda^2}{\kappa^2[10, |\theta_0|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[10, |\theta_0|_0, \sqrt{\Sigma^{(2)}}]}|\theta_0|_0\,. \quad \text{(A.33)}$$

$$|\widehat{\theta}_\lambda|_0 \leq 2^{12}\frac{\Phi_{k_*,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{k_*,+}(\sqrt{\Sigma^{(2)}})}{\kappa^2[10, |\theta_0|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[10, |\theta_0|_0, \sqrt{\Sigma^{(2)}}]}|\theta_0|_0 \leq k_*/2\,, \quad \text{(A.34)}$$

In the sequel, we fix

$$\lambda = 16(\sigma_1 \vee \sigma_2)\sqrt{2\left[\Phi_{1,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{1,+}(\sqrt{\Sigma^{(2)}})\right]\log(p)}\,.$$

and we consider the set $\widehat{S}_\lambda = \mathrm{supp}(\widehat{\theta}^{(1)}_\lambda) \cup \mathrm{supp}(\widehat{\theta}^{(2)}_\lambda)$. On the event $\mathcal{A} \cap \mathcal{B}$, this set $\widehat{S}_\lambda$ belongs to $\widehat{\mathbf{S}}_{\text{Lasso}}$ and its size is smaller or equal to $k_*$ by Lemma A.12. We shall prove that

$$\min_{i\in\{1,2,3\}} \widetilde{Q}_{i,|\widehat{S}_\lambda|}\left(F_{\widehat{S}_\lambda,i}\middle|\mathbf{X}_{\widehat{S}_\lambda}\right) < \alpha_{\widehat{S}_\lambda}$$

with probability larger than $1 - \delta/2$. In the following lemma, we relate the $\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda)$ to $\mathcal{K}_1 + \mathcal{K}_2$.

**Lemma A.13** *On the event* $\mathcal{A} \cap \mathcal{B}$,

$$\begin{aligned}\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda) \;\geq\; & \frac{1}{12}\left[\mathcal{K}_1 + \mathcal{K}_2\right] \\ & -2^{19}\frac{\bigvee_{i=1,2}\Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2}\Phi_{k_*,-}(\sqrt{\Sigma^{(i)}})}\frac{\bigvee_{i=1,2}\Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2}\kappa^2[10, |\theta_0|_0, \sqrt{\Sigma^{(i)}}]}\frac{|S^{\cup(1,2)}|}{n_1 \wedge n_2}\log(p)\,.\end{aligned}$$

Then, we closely follow the arguments of Theorem 4.7 to state that $T^B_\alpha$ rejects $\mathbf{H}_0$ with large probability as long as $\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda)$.

**Lemma A.14** *If on the event* $\mathcal{A} \cap \mathcal{B}$,

$$\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda) \geq 2^{22}\varphi_{\widehat{S}_\lambda}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left[15|\widehat{S}_\lambda|\log(p) + 7\log\{6/(\alpha\delta)\} + 2\log(p)\right]\,,$$

*then,* $\min_{i\in\{1,2,3\}} \widetilde{Q}_{i,|\widehat{S}_\lambda|}(F_{\widehat{S}_\lambda,i}|\mathbf{X}_{\widehat{S}_\lambda}) < \alpha_{\widehat{S}_\lambda}$ *with probability larger than* $1 - \delta/2$.

We derive from (A.34) that on the event $\mathcal{A} \cap \mathcal{B}$,

$$|\widehat{S}_\lambda| \leq 2^9\frac{\bigvee_{i=1,2}\Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2}\kappa^2[10, |\theta_0|_0, \sqrt{\Sigma^{(i)}}]}|S^{\cup(1,2)}|_0\,,$$

Gathering Lemmas A.13 and A.14 allows us to conclude.

*Lemma A.11.* In order to bound $\mathbb{P}(\mathcal{A})$, we apply Lemma A.16 to simultaneously control $\varphi_{\max}(\mathbf{X}_S^{(1*)}\mathbf{X}_S^{(1)})$, $\varphi_{\max}(\mathbf{X}_S^{(2*)}\mathbf{X}_S^{(2)})$, $\varphi_{\min}(\mathbf{X}_S^{(1*)}\mathbf{X}_S^{(1)})$, and $\varphi_{\min}(\mathbf{X}_S^{(2*)}\mathbf{X}_S^{(2)})$ for all sets $S$ of size $k_*$. Combining an union bound with Conditions (A.29) and (A.30) allows us to prove that

$$\mathbb{P}\left[\left\{\forall\theta \text{ s.t. } |\theta|_0 \leq k_*,\ 1/2 \leq \frac{\|\mathbf{X}^{(1)}\theta\|^2}{n_1\|\theta\|^2_{\Sigma^{(1)}}} \leq 2 \text{ and } 1/2 \leq \frac{\|\mathbf{X}^{(2)}\theta\|^2}{n_2\|\theta\|^2_{\Sigma^{(2)}}} \leq 2\right\}\right] \geq 1-\delta/6$$

Applying Corollary 1 in Raskutti et al. (2010), we derive that there exist three positive constant $c_1$, $c_2$ and $c_3$ such that the following holds. With probability larger than $1 - c_1\exp[-c_2(n_1 \wedge n_2)]$, we have

$$\bigwedge_{i=1,2} \frac{\kappa\left[10, |\beta^{(1)}|_0 + |\beta^{(2)}|_0, \mathbf{X}^{(i)}/\sqrt{n_i}\right]}{\kappa\left[10, |\beta^{(1)}|_0 + |\beta^{(2)}|_0, \sqrt{\Sigma^{(i)}}\right]} \geq 2^{-3}\,,$$

if $\left(|\beta^{(1)}|_0 + |\beta^{(2)}|_0\right)\log(p) < c_3 \frac{\vee_{i=1,2}\Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\wedge_{i=1,2}\kappa^2\left[10,|\beta^{(1)}|_0+|\beta^{(2)}|_0,\sqrt{\Sigma^{(i)}}\right]}(n_1 \wedge n_2)$. Hence, we conclude that $\mathbb{P}[\mathcal{A}] \geq 1 - \delta/3$.

Consider an integer $k \leq k_*$ and $\theta$ a $k$-sparse vector. Under the event $\mathcal{A}$, we have

$$\begin{aligned}
\|\mathbf{W}\theta\|^2 &= \|\mathbf{X}^{(1)}/\sqrt{n_1}(\theta^{(1)} + \theta^{(2)})\|^2 + \|\mathbf{X}^{(2)}/\sqrt{n_2}(\theta^{(1)} - \theta^{(2)})\|^2 \\
&\leq 2\|(\theta^{(1)} + \theta^{(2)}\|^2_{\Sigma^{(1)}} + 2\|\theta^{(1)} - \theta^{(2)}\|^2_{\Sigma^{(2)}} \\
&\leq 4\left[\Phi_{k,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{k,+}(\sqrt{\Sigma^{(2)}})\right]\|\theta\|^2 \\
\|\mathbf{W}\theta\|^2 &\geq \frac{1}{2}\left[\|(\theta^{(1)} + \theta^{(2)}\|^2_{\Sigma^{(1)}} + \|\theta^{(1)} - \theta^{(2)}\|^2_{\Sigma^{(2)}}\right] \\
&\geq \left[\Phi_{k,-}(\sqrt{\Sigma^{(1)}}) \wedge \Phi_{k,-}(\sqrt{\Sigma^{(2)}})\right]
\end{aligned}$$

□

*Lemma A.12.* A slight variation of Theorem 14 in Koltchinski et al. (2011) ensures that

$$\|\mathbf{W}(\widehat{\theta}_\lambda - \theta_0)\|^2 \leq \frac{\lambda^2}{\kappa^2[5, |\theta_0|_0, \mathbf{W}]}|\theta_0|_0 \tag{A.35}$$

on the event $\mathcal{B}$. Consider a set $T \subset \{1, \ldots, 2p\}$ of size smaller or equal to $k$ and define $T' \subset \{1, \ldots, p\}$ by $i \in T'$ if $i \in T$ or $i + p \in T$. Consider some

$$\theta = \begin{pmatrix} \theta^{(1)} \\ \theta^{(2)} \end{pmatrix} \in \mathcal{C}(5, T)\,,$$

then either $\theta^{(1)} + \theta^{(2)} \in \mathcal{C}(10, T')$ or $\theta^{(1)} - \theta^{(2)} \in \mathcal{C}(10, T')$. Hence,

$$\begin{aligned}
\frac{\|\mathbf{W}\theta\|^2}{\|\theta\|^2} &= \frac{\|\mathbf{X}^{(1)}(\theta^{(2)} + \theta^{(1)})\|^2}{n_1\|\theta\|^2} + \frac{\|\mathbf{X}^{(2)}(\theta^{(2)} - \theta^{(1)})\|^2}{n_2\|\theta\|^2} \\
&\geq \frac{\|\theta^{(2)} + \theta^{(1)}\|^2 \vee \|\theta^{(2)} - \theta^{(1)}\|^2}{\|\theta\|^2}\left[\bigwedge_{i=1,2}\kappa^2\left(10, k, \mathbf{X}^{(i)}/\sqrt{n_i}\right)\right] \\
&\geq \left[\kappa^2\left(10, k, \mathbf{X}^{(1)}/\sqrt{n_1}\right) \wedge \kappa^2\left(10, k, \mathbf{X}^{(2)}/\sqrt{n_2}\right)\right] \\
&\geq 2^{-6}\left[\kappa^2\left(10, k, \sqrt{\Sigma^{(1)}}\right) \wedge \kappa^2\left(10, k, \sqrt{\Sigma^{(2)}}\right)\right]\,,
\end{aligned}$$

where the last inequality proceeds from Lemma A.11. Hence,

$$2^6\kappa^2[5,|\theta_0|_0,\mathbf{W}] \geq \left[\kappa^2\left(10,k,\sqrt{\Sigma^{(1)}}\right)\wedge\kappa^2\left(10,k,\sqrt{\Sigma^{(2)}}\right)\right] .$$

Gathering this bound with (A.35), it follows that

$$\|\mathbf{W}(\widehat{\theta}_\lambda-\theta_0)\|^2 \leq \frac{2^6\lambda^2}{\kappa^2[10,|\theta_0|_0,\sqrt{\Sigma^{(1)}}]\wedge\kappa^2[10,|\theta_0|_0,\sqrt{\Sigma^{(2)}}]}|\theta_0|_0 ,$$

which allows us to prove (A.33). Lemma 3.1 in Giraud et al. (2012) tells us that on the event $\mathcal{B}$,

$$\lambda^2|\widehat{\theta}_\lambda|_0 \leq 16\Phi_{|\widehat{\theta}_\lambda|_0,+}(\mathbf{W})\|\mathbf{W}(\widehat{\theta}_\lambda-\theta_0)\|^2 .$$

Gathering the two last bounds and Lemma A.11, we obtain

$$|\widehat{\theta}_\lambda|_0 \leq 2^{10}\frac{\Phi_{|\widehat{\theta}_\lambda|_0,+}(\mathbf{W})}{\kappa^2[10,|\theta_0|_0,\sqrt{\Sigma^{(1)}}]\wedge\kappa^2[10,|\theta_0|_0,\sqrt{\Sigma^{(2)}}]}\|\theta_0\|_0.$$

The upper-bound $\Phi_{|\widehat{\theta}_\lambda|_0,+}(\mathbf{W}) \leq (1+\|\widehat{\theta}_\lambda\|_0/k_*)\Phi_{k_*,+}(\mathbf{W})$ and Lemma A.11 enforce

$$\begin{aligned}|\widehat{\theta}_\lambda|_0 &\leq 2^{12}\frac{\Phi_{k_*,+}(\sqrt{\Sigma^{(1)}})\vee\Phi_{k_*,+}(\sqrt{\Sigma^{(2)}})}{\kappa^2[10,|\theta_0|_0,\sqrt{\Sigma^{(1)}}]\wedge\kappa^2[10,|\theta_0|_0,\sqrt{\Sigma^{(2)}}]}|\theta_0|_0\left[1+\frac{|\widehat{\theta}_\lambda|_0}{k_*}\right] \quad \text{(A.36)}\\ &\leq \left(k_*+|\widehat{\theta}_\lambda|_0\right)/2 ,\end{aligned}$$

where the last inequality follows from (A.31) and $|\theta_0|_0 \leq |\beta^{(1)}|_0+|\beta^{(2)}|_0$. Hence, $|\widehat{\theta}_\lambda|_0 \leq k_*$. Coming back to (A.36), we prove (A.34).

□

*Lemma A.13.* Given the Lasso estimator $\widehat{\theta}_\lambda$ of $\theta_0$ in the model (A.32), we define $\widehat{\beta}^{(1)}_\lambda$ and $\widehat{\beta}^{(2)}_\lambda$ by

$$\widehat{\beta}^{(1)}_\lambda = \frac{\widehat{\theta}^{(1)}_\lambda+\widehat{\theta}^{(2)}_\lambda}{\sqrt{n_1}} , \qquad \widehat{\beta}^{(2)}_\lambda = \frac{\widehat{\theta}^{(1)}_\lambda-\widehat{\theta}^{(2)}_\lambda}{\sqrt{n_2}}.$$

On the event $\mathcal{A}\cap\mathcal{B}$, we upper bound the difference between $\beta^{(1)}$ and $\widehat{\beta}^{(1)}_\lambda$ and $\beta^{(2)}$ and $\widehat{\beta}^{(2)}_\lambda$

$$\begin{aligned}&\|\beta^{(1)}-\widehat{\beta}^{(1)}_\lambda\|^2_{\Sigma_1}+\|\beta^{(2)}-\widehat{\beta}^{(2)}_\lambda\|^2_{\Sigma_2}\\ &\leq 2\left[\|\frac{\mathbf{X}^{(1)}}{\sqrt{n_1}}(\beta^{(1)}-\widehat{\beta}^{(1)}\|^2+\|\frac{\mathbf{X}^{(2)}}{\sqrt{n_2}}(\beta^{(2)}-\widehat{\beta}^{(2)}\|^2\right]\\ &\leq \frac{2}{n_1\wedge n_2}\|\mathbf{W}(\theta_0-\widehat{\theta}_\lambda)\|^2\\ &\leq 2^{17}\frac{\bigvee_{i=1,2}\Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2}\kappa^2[10,|\theta_0|_0,\sqrt{\Sigma^{(i)}}]}\frac{|S^{\cup(1,2)}|}{n_1\wedge n_2}\log(p)(\sigma^{(1)}\vee\sigma^{(2)})^2 ,\end{aligned}$$

where the last inequality follows from Lemma A.12. Let us now lower bound the Kullback discrepansy $2\left[\mathcal{K}_1\left(\widehat{S}_\lambda\right)+\mathcal{K}_2\left(\widehat{S}_\lambda\right)\right]$ which equals

$$\frac{(\sigma^{(1)}_{\widehat{S}_\lambda})^2}{(\sigma^{(2)}_{\widehat{S}_\lambda})^2}+\frac{(\sigma^{(1)}_{\widehat{S}_\lambda})^2}{(\sigma^{(2)}_{\widehat{S}_\lambda})^2}-2+\frac{\|\beta^{(2)}_{\widehat{S}_\lambda}-\beta^{(1)}_{\widehat{S}_\lambda}\|^2_{\Sigma^{(2)}}}{(\sigma^{(1)}_{\widehat{S}_\lambda})^2}+\frac{\|\beta^{(2)}_{\widehat{S}_\lambda}-\beta^{(1)}_{\widehat{S}_\lambda}\|^2_{\Sigma^{(1)}}}{\sigma^2_{2,\widehat{S}_\lambda}} \ ?$$

**CASE 1**: $\frac{(\sigma^{(1)}\vee\sigma^{(2)})^2}{(\sigma^{(1)}\wedge\sigma^{(2)})^2}\geq 2$. By symmetry, we can assume that $\sigma^{(1)}>\sigma^{(2)}$

$$\begin{aligned}
(\sigma^{(1)}_{\widehat{S}_\lambda})^2 &= (\sigma^{(1)})^2+\|\beta^{(1)}-\beta^{(1)}_{\widehat{S}_\lambda}\|^2_{\Sigma^{(1)}}\geq(\sigma^{(1)})^2\\
(\sigma^{(2)}_{\widehat{S}_\lambda})^2 &= (\sigma^{(2)})^2+\|\beta^{(1)}-\widehat{\beta}^{(1)}_{\widehat{S}_\lambda}\|^2_{\Sigma^{(1)}}\\
&\leq (\sigma^{(2)})^2+2^{17}\frac{\bigvee_{i=1,2}\Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2}\kappa^2[10,|\theta_0|_0,\sqrt{\Sigma^{(i)}}]}\frac{|S^{\cup(1,2)}|}{n_1\wedge n_2}\log(p)(\sigma^{(1)}\vee\sigma^{(2)})^2\\
&\leq (\sigma^{(2)})^2+\frac{(\sigma^{(1)})^2}{4}\,,
\end{aligned}$$

where we used conditions (A.29) and (A.31) in the last inequality. This enforces

$$2\left[\mathcal{K}_1\left(\widehat{S}_\lambda\right)+\mathcal{K}_2\left(\widehat{S}_\lambda\right)\right]\geq\frac{1}{12}$$

**CASE 2**: $\frac{(\sigma^{(1)}\vee\sigma^{(2)})^2}{(\sigma^{(1)}\wedge\sigma^{(2)})^2}<2$. Let us note

$$A=2^{18}\frac{\bigvee_{i=1,2}\Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2}\kappa^2[10,|\theta_0|_0,\sqrt{\Sigma^{(i)}}]}\frac{|S^{\cup(1,2)}|}{n_1\wedge n_2}\log(p)$$

Arguing as in CASE 1, we derive that

$$\begin{aligned}
(\sigma^{(1)}_{\widehat{S}_\lambda})^2&\leq(\sigma^{(1)})^2[1+A]\leq 2\sigma_1^2\,,\\
\sigma^2_{2,\widehat{S}_\lambda}&\leq(\sigma^{(2)})^2[1+A]\leq 2\sigma_2^2\,.
\end{aligned}$$

Let us lower bound $\mathcal{K}_1(\widehat{S}_\lambda)+\mathcal{K}_2(\widehat{S}_\lambda)$ in terms of $\mathcal{K}_1+\mathcal{K}_2$. First, we consider the ratio of the variances

$$\begin{aligned}
\frac{(\sigma^{(1)}_{\widehat{S}_\lambda})^2}{\sigma^2_{2,\widehat{S}_\lambda}}+\frac{\sigma^2_{2,\widehat{S}_\lambda}}{(\sigma^{(1)}_{\widehat{S}_\lambda})^2}-2 &\geq \left[\frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2}+\frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2}\right]/(1+A)-2\\
&\geq \frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2}+\frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2}-2-\frac{A}{1+A}\left[\frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2}+\frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2}\right]\\
&\geq \frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2}+\frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2}-2-3A\,. \qquad \text{(A.37)}
\end{aligned}$$

Let us now lower bound the remaining part of $\mathcal{K}_1(\widehat{S}_\lambda)+\mathcal{K}_2(\widehat{S}_\lambda)$ For $i=1,2$, the number of non zero components of $\beta^{(i)}-\widehat{\beta}^{(i)}_\lambda$ is smaller or equal

to $k_*$.

$$\begin{aligned}
&\frac{\|\beta^{(1)}-\beta^{(2)}\|^2_{\Sigma^{(2)}}}{(\sigma^{(1)})^2}+\frac{\|\beta^{(1)}-\beta^{(2)}\|^2_{\Sigma^{(1)}}}{(\sigma^{(2)})^2}\\
&\quad\leq\quad \frac{3}{(\sigma^{(1)})^2\wedge\sigma_2^2}\sum_{i=1}^{2}\left[\|\beta^{(1)}-\beta^{(1)}_{\widehat{S}_\lambda}\|^2_{\Sigma^{(i)}}+\|\beta^{(2)}-\beta^{(2)}_{\widehat{S}_\lambda}\|^2_{\Sigma^{(i)}}+\|\beta^{(1)}_{\widehat{S}_\lambda}-\beta^{(2)}_{\widehat{S}_\lambda}\|^2_{\Sigma^{(i)}}\right]\\
&\quad\leq\quad \frac{3}{(\sigma^{(1)})^2\wedge\sigma_2^2}\left[\left[1+\frac{\Phi_{k_*,+}(\sqrt{\Sigma^{(1)}})}{\Phi_{k_*,-}(\sqrt{\Sigma^{(2)}})}\right]\sum_{i=1}^{2}\|\beta^{(i)}-\widehat{\beta}^{(i)}_\lambda\|^2_{\Sigma^{(i)}}+\sum_{i=1}^{2}\|\beta^{(1)}_{\widehat{S}_\lambda}-\beta^{(2)}_{\widehat{S}_\lambda}\|^2_{\Sigma^{(i)}}\right]\\
&\quad\leq\quad 12\left[\sum_{i=1}^{2}\frac{\|\beta^{(1)}_{\widehat{S}_\lambda}-\beta^{(2)}_{\widehat{S}_\lambda}\|^2_{\Sigma^{(i)}}}{\sigma^2_{(i+1)\bmod 2}}\right]+2^3\frac{\bigvee_{i=1,2}\Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2}\Phi_{k_*,-}(\sqrt{\Sigma^{(i)}})}A
\end{aligned}$$

Gathering the last inequality with (A.37) yields

$$\mathcal{K}_1(\widehat{S}_\lambda)+\mathcal{K}_2(\widehat{S}_\lambda)\geq\frac{1}{12}\left[\mathcal{K}_1+\mathcal{K}_2\right]-3\frac{\bigvee_{i=1,2}\Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2}\Phi_{k_*,-}(\sqrt{\Sigma^{(i)}})}A\,.$$

□

*Lemma A.14.* For any non empty set $S$ of size smaller or equal to $k_*$, define $\delta_S=\delta\left(2\binom{|S|}{p}k_*\right)^{-1}$. By definition (A.29) of $k_*$ and by Hypothesis (A.30), the following conditions are satisfied

$$\log(12/\delta_S)<2^{-13}(n_1\wedge n_2),\quad \log(1/\alpha_S)\leq 2^{-10}(n_1\wedge n_2)\,.$$

Arguing as in the proof of Theorem 4.7, we have

$$\mathbb{P}\left[\min_{i\in\{1,2,3\}}\widetilde{Q}_{i,|\widehat{S}_\lambda|}(F_{\widehat{S}_\lambda,i}|\mathbf{X}_{\widehat{S}_\lambda})<\alpha_{\widehat{S}_\lambda}\right]\geq 1-\delta_S$$

as long as

$$\mathcal{K}_1(\widehat{S}_\lambda)+\mathcal{K}_2(\widehat{S}_\lambda)\geq 2^{22}\varphi_{\widehat{S}_\lambda}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left[15|\widehat{S}_\lambda|\log(p)+7\log\{6/(\alpha\delta)\}+2\log(p)\right],$$

Applying an union bound over all sets $S$ of size smaller or equal to $k_*$ allows us to conclude.

□

### A.2.5 Technical lemmas

In this section, some useful deviation inequalities for $\chi^2$ random variables Laurent and Massart (2000) and for Wishart matrices Davidson and Szarek (2001) are reminded.

**Lemma A.15** *For any integer $d>0$ and any positive number $x$,*

$$\begin{aligned}
\mathbb{P}\left(\chi^2(d)\leq d-2\sqrt{dx}\right)&\leq\exp(-x)\,,\\
\mathbb{P}\left(\chi^2(d)\geq d+2\sqrt{dx}+2x\right)&\leq\exp(-x)\,.
\end{aligned}$$

**Lemma A.16** *Let $Z^\intercal Z$ be a standard Wishart matrix of parameters $(n, d)$ with $n > d$. For any positive number $x$,*

$$\mathbb{P}\left\{\varphi_{\min}\left(Z^\intercal Z\right) \geq n\left(\left\{1-\sqrt{\frac{d}{n}}-x\right\}\vee 0\right)\right\} \leq \exp(-nx^2/2)\,,$$

*and*

$$\mathbb{P}\left[\varphi_{\max}\left(Z^\intercal Z\right) \leq n\left(1+\sqrt{\frac{d}{n}}+x\right)^2\right] \leq \exp(-nx^2/2)\,.$$

# Bibliography


S. Allassonnière and C. Giraud. Detecting long distance conditional correlations between anatomical regions using Gaussian graphical models. In *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy - Geometrical and Statistical Methods for Modelling Biological Shape Variability*, pages 111–122, 2011. (Cité page 30.)

C. Ambroise, J. Chiquet, and C. Matias. Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238, 2009. (Cité pages 3, 28, 30, 39, 40, and 42.)

A. Anandkumar, V. Tan, and A. Willsky. High-dimensional graphical model selection: Tractable graph families and necessary conditions. In *NIPS*, 2011. (Cité page 122.)

A. Antoniadis. Comments on: l1-penalization for mixture regression models. *Test*, 19:257–258, 2010. (Cité page 89.)

E. Arias-Castro, E. Candes, and Y. Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Annals of Statistics*, 39:2533–2556, 2011. (Cité pages 90, 103, 104, and 105.)

F. Bach. Bolasso: model consistent Lasso estimation through the bootstrap. In *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*, 2008a. (Cité page 30.)

F. Bach. Consistency of the group-Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008b. (Cité pages 61, 75, and 126.)

F. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, 2010. (Cité page 30.)

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4:1–106, 2012. (Cité pages 30 and 72.)

Z. Bai and H. Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6:311–329, 1996. (Cité page 88.)

S. Bakin. *Adaptive regression and model selection in data-mining problems*. PhD thesis, Australian National University, Canberra, 1999. (Cité page 60.)

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008. (Cité pages 27 and 39.)

Y. Baraud, S. Huet, and B Laurent. Adaptive tests of linear hypotheses by model selection. *Annals of Statistics*, 31:225–251, 2003. (Cité pages 90, 93, 102, and 135.)

Y. Baraud, C. Giraud, and S. Huet. Estimator selection in the gaussian setting. Technical report, arXiv, 2010. (Cité pages 25 and 110.)

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009. (Cité pages 48, 87, and 129.)

J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer-Verlag, 2006. (Cité page 72.)

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, third edition, 2006. (Cité page 72.)

P. Bühlmann. Causal statistical inference in high dimensions. *Mathematical Methods of Operations Research*, to appear, 2012a. (Cité page 122.)

P. Bühlmann. Statistical significance in high-dimensional linear models. Technical report, arXiv, 2012b. (Cité pages 89 and 90.)

T. Cai, J. Jin, and M. Low. Estimation and confidence sets for sparse normal mixtures. *Annals of Statistics*, 35:2421–2449, 2007. (Cité page 104.)

E. Candes and Y. Plan. Near-ideal model selection by $\ell_1$ minimization. *Annals of Statistics*, 37:2145–2177, 2007. (Cité page 103.)

R. Castelo and A. Roverato. A robust procedure for Gaussian graphical model search from microarray data with $p$ larger than $n$. *Journal of Machine Learning Research*, 7:2621–2650, 2006. (Cité page 27.)

C. Charbonnier, J. Chiquet, and C. Ambroise. Weighted-Lasso for structured network inference from time-course data. *Statistical Applications in Genetics and Molecular Biology*, 9(1):15, 2010. (Cité pages 33 and 45.)

J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771, 2008. (Cité page 25.)

S. Chen and D. Donoho. Basis pursuit. In *28th Asilomar Conference on Signals, Systems and Computers*, 1994. (Cité page 15.)

S. Chen and Y. Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics*, 38:808–835, 2010. (Cité page 88.)

J. Chiquet, Y. Grandvalet, and C. Ambroise. Inferring multiple graphical structures. *Statistics and Computing*, 21(4):537–553, 2011. (Cité pages 3, 57, 59, 60, 63, 72, 79, and 82.)

J. Chiquet, Y. Grandvalet, and C. Charbonnier. Sparsity with sign-coherent groups of variables via the cooperative-Lasso. *The Annals of Applied Statistics*, to appear, 2012. (Cité pages 57 and 61.)

J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, 2008. (Cité page 44.)

K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*, pages 317–366. North-Holland, Amsterdam, 2001. (Cité page 146.)

A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212, 2004. (Cité page 27.)

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixture. *Annals of Statistics*, 32:962–994, 2004. (Cité pages 90, 104, and 105.)

D. Donoho and J. Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367:4449 – 4470, 2009. (Cité page 104.)

D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52:6–18, 2006. (Cité page 21.)

C. Dossal, M. Kachour, J. Fadili, G. Peyré, and C. Chesneau. The degrees of freedom of the lasso in underdetermined linear regression models. In *SPARS 11, Edinburg*, 2011. (Cité page 25.)

M. Drton and M. Perlman. Multiple testing and error control in Gaussian graphical model selection. *Statististical Science. A Review Journal of the Institute of Mathematical Statistics*, 22:430, 2007. (Cité page 27.)

M. Drton and M. Perlman. A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138:1179–1200, 2008. (Cité page 27.)

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors. (Cité pages 24 and 108.)

M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 95 (25):14863–8., 1998. (Cité pages 59 and 61.)

R. Foygel and M. Drton. Extended bayesian information criteria for gaussian graphical models. In *NIPS*, 2010. (Cité page 25.)

O. Frank and F. Harary. Cluster inference by using transitivity indices in empirical graphs. *Journal of the American Statistical Association*, 77(380): 835–840, 1982. (Cité page 44.)

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, 2008. (Cité page 27.)

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group-Lasso and a sparse-group-Lasso. Technical report, arXiv, 2010. (Cité page 60.)

X. Gao, D. Pu, Y. Wu, and H. Xu. Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statistica Sinica*, 22: 1123–1146, 2012. (Cité page 25.)

C. Giraud, S. Huet, and N. Verzelen. Supplement to 'High-dimensional regression with unknown variance', 2012. (Cité pages 142 and 144.)

Y. Grandvalet and S. Canu. Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In *Advances in Neural Information Processing Systems 11 (NIPS 1998)*, pages 445–451, 1999. (Cité page 60.)

R. Gribonval. Should penalized least squares regression be interpreted as maximum a posteriori estimation? *IEEE Transactions on Signal Processing*, 59:2405–2410, 2011. (Cité page 40.)

M. Guedj, L Marisa, A. de Reynies, B. Orsetti, N. Schiappa, F. Bibeau, G. MacGrogan, F. Lerebours, P. Finetti, M. Longy, P. Bertheau, F. Bertrand, F. Bonnet, AL Martin, JP Feugeas, I. Bièche, J. Lehmann-Che, R. Lidereau, D. Birnbaum, F. Bertucci, H. de The, and C. Theillet. A refined molecular taxonomy of breast cancer. *Oncogene*, 31:1196–1206, 2012. (Cité page 83.)

P. Hall and J. Jin. Properties of higher criticism under strong dependence. *Annals of Statistics*, 36:381–402, 2008. (Cité page 90.)

P. Hall and J. Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *Annals of Statistics*, 38:1686–1732, 2010. (Cité page 105.)

J. Haupt, R. Castro, and R. Nowak. Adaptive discovery of sparse signals in noise. In *42th Asilomar Conference on Signal, Systems and Computers, Pacific Grove, California*, 2008. (Cité page 104.)

J. Haupt, R. Castro, and R. Nowak. Distilled sensing: Adap- tive sampling for sparse detection and estimation. Technical report, arXiv:1001.5311, 2010. (Cité page 104.)

A-C Haury, P Gestraud, and J-P Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 6, 2011a. (Cité page 1.)

A-C Haury, F Mordelet, P Vera-Licona, and J-P Vert. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *MLCB (NIPS workshop)*, 2011b. (Cité page 30.)

A. Hauser and P. Bühlmann. Two optimal strategies for active learning of causal models from interventions. Technical report, arXiv, 2012. (Cité page 122.)

J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Statistics*, 38:1978–2004, 2010. (Cité page 60.)

Y. Ingster, A. Tsybakov, and N. Verzelen. Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4:1476–1526, 2010. (Cité pages 90, 103, 104, and 105.)

J Ioannidis. Microarrays and molecular research: noise discovery? *Lancet*, 365, 2005. (Cité page 1.)

A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *NIPS*, pages 964–972, 2010. (Cité pages 60 and 62.)

A. Jalali, P. Ravikumar, and S. Sanghavi. A dirty model for multiple sparse regression. *CoRR*, abs/1106.5826, 2011. (Cité pages 60, 62, and 126.)

M Jeanmougin. Improving gene signatures by the identification of differentially expressed modules in molecular networks : a local-score approach. In *JOBIM Rennes*, 2012. (Cité page 83.)

M. Jeanmougin, M. Guedj, and C. Ambroise. Defining a robust biological prior from pathway analysis to drive network inference. *Journal de la Société Française de Statistique*, 152:97–110, 2011. (Cité page 42.)

B. Jones, C. Carvalho, A. Dobra, C. Hans, C. Carter, and M. West. Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, 20(4):388–400, 2005. (Cité page 27.)

A Juditsky and A. Nemirovsky. On verifiable sufficient conditions for sparse signal recovery via $\ell_1$ minimization. *ArXiv*, 2008. URL `http://arxiv.org/abs/0809.2650`. (Cité page 48.)

Harri Kiiveri. Multivariate analysis of microarray data: differential expression and differential connection. *BMC Bioinformatics*, 12(1):42, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-42. URL `http://www.biomedcentral.com/1471-2105/12/42`. (Cité page 27.)

K. Knight and W. Fu. Asymptotics for Lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000. (Cité pages 18, 39, and 127.)

V. Koltchinski, K. Lounici, and A. Tsybakov. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011. (Cité pages 142 and 143.)

J Kovalchin, J Krieger, M. Genova, K. Collins, M Augustyniak, A. Masci, T. Hittinger, B. Kuca, G. Edan, C. Braudeau, M. Rimbert, U. Patel, E. Mascioli, and E. Zanelli. Results of a phase i study in patients suffering from secondary-progressive multiple sclerosis demonstrating the safety of the amino acid copolymer pi-2301 and a possible induction of an anti-inflammatory cytokine response. *Journal of Neuroimmunology*, 225:153–63, 2010. (Cité page 82.)

M. Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27:303–324, 2009. (Cité page 62.)

M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis*, 5:369–412, 2010. (Cité page 89.)

P. Latouche, E. Birmele, and C. Ambroise. Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 2011. URL `http://arxiv.org/abs/0912.2873v2`. (Cité page 44.)

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000. ISSN 0090-5364. (Cité page 146.)

S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996. (Cité pages 7 and 8.)

S. Lèbre. Inferring dynamic genetic networks with low order independencies. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009. (Cité pages 35 and 46.)

S. Lèbre, J. Becq, F. Devaux, M. P. H. Stumpf, and G. Lelandais. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, 4(130):1–16, 2010. URL `http://www.biomedcentral.com/1752-0509/4/130`. (Cité page 37.)

M. Lopes, L. Jacob, and M. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *NIPS*, 2011. (Cité pages 88 and 90.)

K. Lounici, M. Pontil, A.B. Tsybakov, and S. van de Geer. Sparsity for multi-task learning. In *Conference On Learning Theory*, 2009. (Cité pages 61, 62, and 129.)

K. Lounici, M. Pontil, A.B. Tsybakov, and S. van de Geer. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39: 2164–2204, 2011. (Cité pages 61, 79, and 129.)

S. Ma, X. Song, and J Huang. Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics*, 8(60), 2007. (Cité pages 59 and 61.)

M. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248, 2010. (Cité page 122.)

M. Mariadassou and S. Robin. Uncovering latent structure in valued graphs: a variational approach. Technical Report 10, Statistics for Systems Biology, 2007. (Cité page 44.)

B. Marlin, M. Schmidt, and K. Murphy. Group sparse priors for cavariance estimation. In *Uncertainty in Artificial Intelligence*, 2009. (Cité page 42.)

R. Mazumder and T. Hastie. The graphical Lasso : New insights and alternatives. Technical report, Stanford University, 2011. (Cité page 27.)

L. Meier, S. Van De Geer, and P. BÃ$\frac{1}{4}$hlmann. The group-Lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70:53–71, 2008. (Cité page 61.)

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006. (Cité pages 21, 27, 28, 35, 39, 45, and 48.)

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72:417–473, 2010. (Cité page 30.)

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of statistics*, 37:246–270, 2009. (Cité pages 47 and 87.)

N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104: 1671–1681, 2009. (Cité pages 89, 90, and 98.)

Y. Nardi and A. Rinaldo. On the asymptotic properties of the group-Lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008. (Cité page 61.)

S. Negahban and M. Wainwright. Simultaneous support recovery in high dimensions: Benefits and perils of block $\ell_1/\ell_\infty$-regularization. *IEEE Transactions on Information Theory*, 57:3841–3863, 2011. (Cité pages 60, 61, 62, 76, and 126.)

K. Nowicki and T. Snijders. Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001. (Cité page 44.)

G. Obozinski, M. Wainwright, and M. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39:1–47, 2011. (Cité pages 61, 62, 76, and 126.)

R. Opgen-Rhein and K. Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive model. *BMC Bioinformatics*, 8, 2007. (Cité pages 35, 39, 46, and 47.)

M. Osborne, B. Presnell, and B. Turlach. On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000. (Cité pages 35 and 89.)

M. Park, T. Hastie, and R. Tibshirani. Averaged gene expressions for regression. *Biostatistics*, 8(2):212–227, 2006. (Cité pages 59 and 61.)

G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11: 2241–2259, 2010. ISSN 1532-4435. (Cité pages 140 and 143.)

A. Rau, F. Jaffrézic, J.-L. Foulley, and R. W. Doerge. Reverse engineering gene regulatory networks using approximate Bayesian computation. Technical report, ArXiv, 2011. URL `http://arXiv.org/abs/1109.1402v1`. (Cité page 27.)

P. Ravikumar, M. Wainwright, and J. Lafferty. High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. *Annals of Statistics*, 38:1287–1319, 2010. (Cité page 28.)

P. Ravikumar, M. Wainwright, G Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011. (Cité page 29.)

G. V. Rocha, P. Zhao, and B. Yu. A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE), 2008. (Cité page 28.)

F. Rohart. Multiple hypotheses testing for variable selection. Technical report, INSA Toulouse, 2011. (Cité page 30.)

M. Ronen, R. Rosenberg, B. Shraiman, and U. Alon. Assigning numberes to the arrows: parametrizing a gene regulation network by using accurate expression kinetics. *PNAS*, 99(16):10555–10560, 2002. (Cité page 53.)

M. J. Wainwright S. Negahban, P. Ravikumar and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, to appear, 2012. (Cité pages 24, 30, 61, 79, 129, 131, and 132.)

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005. (Cité page 27.)

T. Shimamura, S. Imoto, R. Yamaguchi, and S. Miyano. Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Informatics*, 19:142 – 153, 2007. (Cité page 35.)

T. Shimamura, S. Imoto, R. Yamaguchi, A. Fujita, M. Nagasaki, and S. Miyano. Recursive regularization for inferring gene networks from time-course gene expression profiles. *BMC Systems Biology*, 3(41), 2009. (Cité pages 35 and 46.)

T. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14 (1):75–100, 1997. (Cité page 44.)

P. Spellman, G. Sherlock, M. Zhang, V. Iyer, M. Eisen, P. Braow, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the cell*, 9:3273–3297, 1998. (Cité pages 47, 49, and 53.)

M. Srivastava and M. Du. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99:386–402, 2008. (Cité page 88.)

T. Sun and C. Zhang. Comments on: l1-penalization for mixture regression models. *Test*, 19:270–275, 2010. (Cité page 89.)

T. Sun and C. Zhang. Scaled sparse linear regression. Technical report, arXiv:1104.4595, 2011. (Cité page 89.)

M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite kernel learning. *Machine Learning Journal*, 79:73–103, 2010. (Cité page 62.)

C. Tallberg. A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29(1):1–23, 2005. (Cité page 44.)

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996. (Cité pages 15, 25, and 89.)

J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52:1030–1051, 2006. (Cité page 21.)

J. Tropp, A. Gilbert, and M. Strauss. Algorithms for simultaneous sparse approximation. *Signal Processing*, 86:572–602, 2006. (Cité page 60.)

B. Turlach, W. Venables, and S. Wright. Simultaneous variable selection. *Technometrics*, 27:349–363, 2005. (Cité pages 59 and 60.)

S. van de Geer. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. (Cité page 21.)

A.W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. (Cité page 126.)

J. Varghese and D. Easton. Genome-wide association studies in common cancers — what have we learnt? *Current Opinion in Genetics and Development*, 20:201–209, 2010. (Cité page 1.)

N. Verzelen. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. *Electron. J. Stat.*, 6:38–90, 2012. (Cité pages 101 and 121.)

N. Verzelen and F. Villers. Tests for gaussian graphical models. *Comput. Statist. Data Anal.*, 53:1894–1905, 2009. (Cité page 114.)

N. Verzelen and F. Villers. Goodness-of-fit tests for high-dimensonal gaussian linear models. *Annals of Statistics*, 38:704–752, 2010. (Cité pages 4, 85, 90, 91, 93, and 101.)

F. Villers, B. Schaeffer, C. Bertin, and S. Huet. Assessing the validity domains of graphical Gaussian models in order to infer relationships among components of complex biological systems. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008. (Cité page 28.)

M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55, 2009a. (Cité pages 22, 87, and 126.)

M. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55, 2009b. (Cité page 121.)

H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B*, 71:671–683, 2009.

W. Wang, M. Wainwright, and K. Ramchandran. Information-theoretic bounds on model selection for gaussian markov random fields. In *IEEE International Symposium on Information Theory, Austin, TX*, 2010. (Cité page 122.)

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of Statistics*, 37:2178–2201, 2009. (Cité pages 89 and 98.)

J. Whittaker. *Graphical models in applied multivariate statistics*. John Wiley and Sons, 1990. (Cité pages 7 and 12.)

A. Wille and P. Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology*, 5(1), 2006. (Cité page 27.)

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 68(1):49–67, 2006. (Cité page 60.)

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007a. (Cité page 27.)

M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of the Royal Statistical Society, Series B*, 69(2):143–161, 2007b. (Cité pages 75 and 126.)

C. Zhang and S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. Technical report, arXiv, 2011. (Cité pages 89 and 90.)

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine learning Research*, 7:2541–2563, 2006. (Cité pages 21 and 47.)

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6): 3468–3497, 2009. (Cité pages 60 and 62.)

S. Zhou, S. van de Geer, and P. Bühlmann. Adaptive Lasso for high dimensional regression and Gaussian graphical modeling. Technical report, ArXiv, 2009. URL `http://arxiv.org/abs/0903.2515v1`. (Cité pages 35, 39, and 45.)

S. Zhou, S. van de Geer, and P. Bühlmann. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, 5:688–749, 2011. URL `http://arxiv.org/abs/0903.2515v1`. (Cité page 30.)

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006. (Cité pages 30, 35, and 39.)

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 67(2):301–320, 2005. (Cité pages 30 and 35.)

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the Lasso. *Annals of Statistics*, 35(5):2173–2192, 2007. (Cité pages 25 and 48.)