

Graphical Model Selection for Gaussian Conditional Random Fields in the Presence of Latent Variables: Theory and Application to Genetics



Benjamin Frot
University College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2016

Acknowledgements

First and foremost, I would like to thank my supervisors Gil McVean and Luke Jostins for their guidance throughout my DPhil. The work presented here would have never been possible without their insightful comments and their patience. All of the ideas contained in this document are the fruit of my conversations with them and of the countless hours they spent listening to my problems and commenting on my work. I had no experience in the fields of statistics and genetics and their rigorous explanations made the learning process infinitely more efficient and pleasant. Beyond the academic support I received, I am also hugely indebted to Gil for his help with many other aspects of my life as a DPhil student. I could hardly overstate how enjoyable my time in Oxford has been and the role played by Luke and Gil in that respect.

I want to thank Jerome Kelleher for his support and his patience. Without him, a lot of the computational aspects of my thesis would have been really challenging, if not impossible.

Finally, I am grateful to all the members of the Department of Statistics who took the time to share their ideas with me and answer my questions. In particular, I would like to thank Professors Robin Evans, Chris Holmes and Yee Whye Teh.

Abstract

The task of performing graphical model selection arises in many applications in science and engineering. The field of application of interest in this thesis relates to the needs of datasets that include genetic and multivariate phenotypic data. There are several factors that make this problem particularly challenging: some of the relevant variables might not be observed, high-dimensionality might cause identifiability issues and, finally, it might be preferable to learn the model over a subset of the collection while conditioning on the rest of the variables, *e.g.* genetic variants. We suggest addressing these problems by learning a conditional Gaussian graphical model, while accounting for latent variables. Building on recent advances in this field, we decompose the parameters of a conditional Markov random field into the sum of a sparse and a low-rank matrix. We derive convergence bounds for this novel estimator, show that it is well-behaved in the high-dimensional regime and describe algorithms that can be used when the number of variables is in the thousands. Through simulations, we illustrate the conditions required for identifiability and show that this approach is consistent in a wider range of settings. In order to show the practical implications of our work, we apply our method to two real datasets and devise a metric that makes use of an independent source of information to assess the biological relevance of the estimates. In our first application, we use the proposed approach to model the levels of 39 metabolic traits conditional on hundreds of genetic variants, in two independent cohorts. We find our results to be better replicated across cohorts than the ones obtained with other methods. In our second application, we look at a high-dimensional gene expression dataset. We find that our method is capable of retrieving as many biologically relevant gene-gene interactions as other methods while retrieving fewer irrelevant interaction.

Contents

1	Introduction	1
1.1	Previous work	7
1.1.1	Graphical models: key concepts and definitions	8
1.1.2	Mendelian randomisation: genetic variants as instrumental variables	10
1.1.2.1	Definition	11
1.1.2.2	Using multiple instruments	13
1.1.2.3	Limitations	15
1.1.3	Inverse covariance estimation	15
1.1.3.1	Problem statement	16
1.1.3.2	Modelling all variables jointly: the graphical lasso	17
1.1.3.3	Conditioning : Sparse Gaussian Conditional Markov Random Fields.	18
1.1.3.4	Marginalising: Low-Rank Plus Sparse Decomposition.	19
1.1.3.5	Limitations	20
1.2	Contributions	21
1.3	Outline	22
2	Problem Statement and Suggested Estimator	24
2.1	Problem statement	24
2.2	Suggested estimator	25
3	Theoretical Results	29
3.1	Key definitions and assumptions	30
3.1.1	Key definitions	31
3.1.2	Conditions on the Fisher Information Matrix (FIM)	32
3.2	Consistency	35

4	Practical Considerations	39
4.1	Fitting	41
4.1.1	Using a proximal gradient algorithm	42
4.1.2	Recasting the objective function as a Semidefinite Program (SDP)	45
4.2	Selecting the tuning parameters	47
4.2.1	Stability selection	48
4.2.2	Sensitivity to γ	52
4.2.3	Suggested procedure	54
4.2.4	Performance on synthetic data	55
4.2.4.1	Data generation	55
4.2.4.2	Results	56
4.3	Assessing the biological relevance of graphical structures	60
4.3.1	Definition	61
4.3.2	Interpretation	62
5	Simulations	65
5.1	Graphical structures	65
5.2	Results	67
6	Application: Using Genetic Information to Detect Relationships between Human Metabolites	72
6.1	The Avon Longitudinal Study of Parents and Children (ALSPAC)	72
6.1.1	Relevance of the ALSPAC dataset	74
6.2	Methods	75
6.2.1	Selection of the tuning parameters	75
6.2.2	Assessing the biological relevance of the estimates	75
6.3	Results	76
6.4	Discussion	82
7	Application: Using Expression Quantitative Trait Loci to Model the Expression Levels of Human Genes	84
7.1	Expression Quantitative Trait Loci (eQTLs)	85
7.2	Data Preparation	87
7.3	Methods and choice of the tuning parameters	89
7.4	Assessing the quality of the estimates	92
7.4.1	General procedure	93
7.4.2	Main measure: the enrichment statistic of Chapter 4	93

7.5	Results	94
7.6	Discussion	99
8	Discussion	105
A	A summary of our notations	114
B	Proof of Theorem 2	117
B.1	Goal	117
B.2	Identifiability	117
B.3	Elementary properties of the likelihood	119
B.4	Curvature of the likelihood and the rank variety	121
B.4.1	Curvature of the rank variety	121
B.4.1.1	Jordan-Wieland Matrix and Matrix perturbation bounds:	122
B.4.1.2	Curvature of the rank variety – Proof of Proposition 2	124
B.4.1.3	Curvature of the likelihood	125
B.5	Consistency	130
B.5.1	Statement of the result	131
B.5.2	Tangent space constraints	133
B.5.3	Variety constraints	138
B.5.4	From variety to tangent space constraints	144
B.5.5	From tangent space constraints to problem (B.1)	145
B.5.6	Bounding the error terms	146
B.5.7	Proof of consistency	151
	Bibliography	154

Chapter 1

Introduction

In recent years, storage costs have been considerably reduced and accessing huge amounts of computational resources has become very easy. The scale and depth of datasets being collected on a daily basis are unprecedented. These advances have led to new opportunities in nearly all industries and areas of science (The Economist 2010). Decision making and discoveries are increasingly based on empirical facts, and not solely on expert opinions. When it comes to making sense of these large amounts of data, the framework of graphical models has proven to be an especially valuable tool (Fan & Liu 2013). Its language makes it possible to reason about complex interactions between variables and enables a principled approach to causal inference (Pearl 2000). These are important features because, in science – and in biology in particular – uncovering causal mechanisms, *i.e.* the data-generating process, is often one of the motivations behind data collection. This places the task of performing graphical model selection from observational data at the heart of many areas of science and engineering, although it would be incorrect to equate causal inference with graphical model estimation.

There are several factors that make graphical model selection particularly challenging. First, it is common that only a subset of the relevant variables are observed so

that is impossible to make claims about the data-generating process without making untestable assumptions (Pearl 2009). Second, the number of variables being modelled is often comparable to or greater than the number of samples. Here again, assumptions (*e.g.* sparsity, low-dimensionality) are necessary in order to obtain consistent estimators (Johnstone 2001). Finally, modelling the joint distribution over all observed variables is not always relevant, as is highlighted by the discriminative/generative contrast in machine learning (Ng & Jordan 2002, Sutton & McCallum 2012). This is the case in genetics, for example, where one could model a gene expression network conditional on the samples' combinations of DNA variants (Yin & Li 2011, Zhang & Kim 2014). Implicit to the idea of conditioning is the desire to encode a notion of direction of effect: the expression of a gene can be influenced by DNA, but no changes in gene expression levels will modify an individual's genome. The result of such a discriminative approach is an increase in power and identifiability because explicit – and valid – assumptions are being made about the data. The need for assumptions being the leitmotiv of the challenges faced by latent graphical model selection, a key task is therefore to develop statistical methods that rely on premises that are 1) easily understood by the researchers who are going to use them; 2) capable of accounting for domain specific knowledge.

A motivating example

The present work is motivated by the need to analyse datasets that include genetic and multivariate phenotypic data. In order to get a better understanding of the kind of problem we are interested in, we give a concrete example. This will give us the opportunity to identify the key features of the sort of dataset we want to extract value from.

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a cohort study of children born in the former county of Avon between 1991 and 1992. ALSPAC

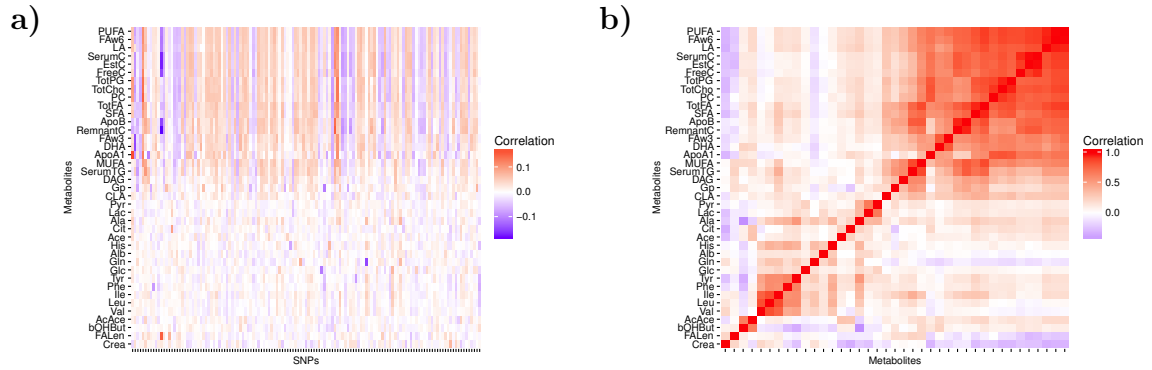


Figure 1.1: a) Correlation matrix between the levels of 39 metabolites and 133 genetic loci. b) Correlation matrix of the levels of 39 metabolites (X)

recruited 14,541 pregnant women resident in the former county of Avon with expected dates of delivery between 1st April 1991 and 31st December 1992. Thousands of the children (and their mothers) enrolled in this study subsequently attended clinics throughout their childhood. In the dataset we will analyse in a later chapter (Chapter 6), genetic and metabolic measurements on more than 5,200 children and 2,700 mothers are available. Some of these metabolic measurements concern the levels of various amino acids, the concentration of $\omega - 3,6$ fatty acids, of triglycerides, of remnant cholesterol etc... We wish to model the levels of these metabolites conditional on genetic data. Metabolites are interesting because they are located far enough from the DNA in the causal chain for the genetics to affect their levels through various pathways. To get a sense of the relationship between genetics and metabolites, consider the covariance matrix plotted in Figure 1.1 **a)**. This figure displays the correlation structure between a number of genetic loci and the levels of 39 metabolites. As can be seen from this plot, it is common for a single locus to affect the levels of dozens of metabolites.

Figure 1.1 **b)** also shows the correlation matrix over metabolites. Given that the levels of many metabolites are strongly correlated, one could expect some of these observed correlations to be the result of some latent variables in which the genetics play

a part, *e.g.* Body Mass Index (BMI). A generic model for this example could therefore be represented by the schema depicted in Figure 1.2. In this diagram, boxed variables are unobserved and red arrows depict the structure of the true graphical model over metabolites. Genetic variants are allowed to act on the observed phenotypes through latent pathways but can also have a direct effect (blue arrows). Furthermore, there is another class of unobserved variables (confounders) that impact observed metabolites only, so that a complex observed correlation structure might effectively be due to a few latent variables and a few direct metabolite-metabolite interactions. In a typical application, one would be interested in inferring the arrows in red while accounting, somehow, for the confounding that might result from the presence of hidden variables and genetic variants. Model 1.2 is very general and is not restricted to metabolites: all datasets that include genetic and multivariate phenotypic data can be described by such a model. For example, it could be used to model low-level phenotypes such as gene expression levels. In this case, the true graph would most likely be such that most of the variants act directly (blue arrows) with few latent pathways and many confounders. We could also find applications outside the realm of biology. For example, it could be well suited to the modelling of the growth rate of the U.S. production indices conditional on hundreds of economic variables (Chen & Huang 2016).

Research questions

Reflecting on this example, we can identify a few distinctive features of model 1.2: 1) due to the presence of latent variables, an infinity of structures are Markov equivalent, thus raising identifiability issues; 2) even if that were not the case, the search space would be so vast that it could not be explored exhaustively; 3) the direction of some arrows is known (*e.g.* from variants to metabolites) while the direction of other arrows is unknown (metabolite/metabolite edges); 4) some variables do not have incoming edges. To the issues raised by these features, we also add the challenges mentioned in

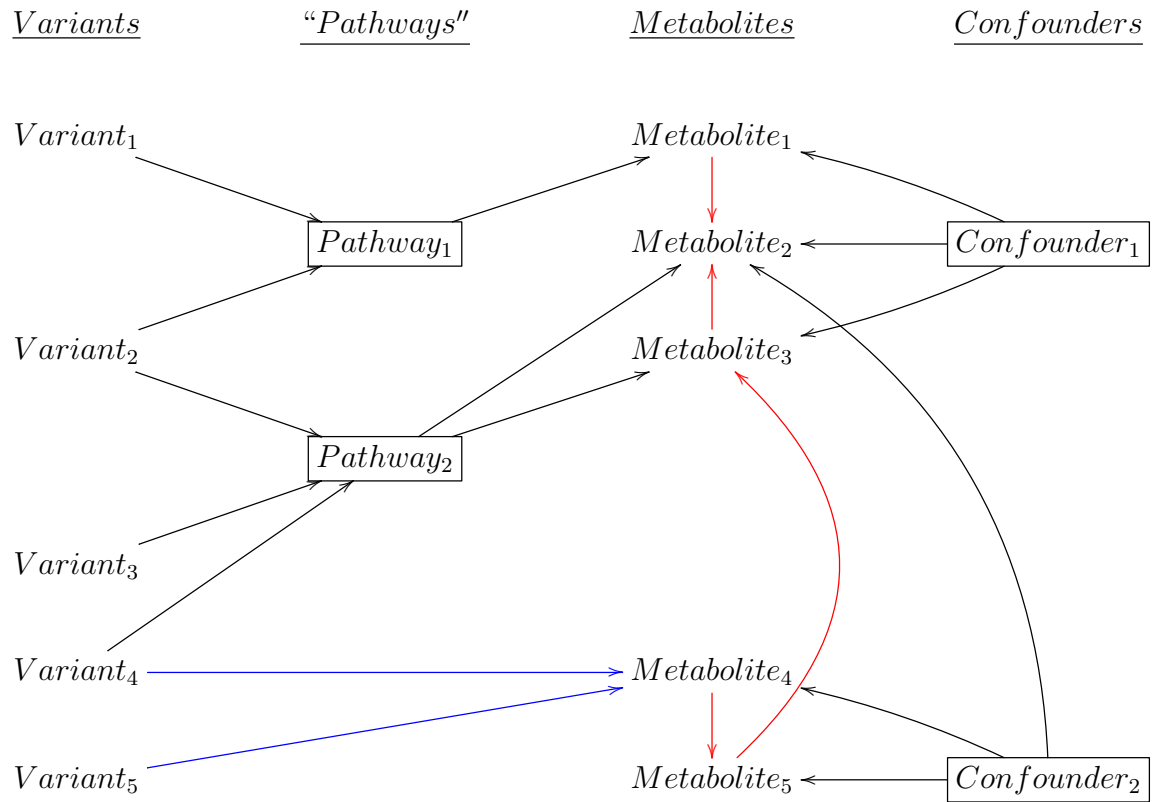


Figure 1.2: A directed graph depicting the relationship between genetic variants, metabolite levels and latent variables. Boxed variables are unobserved; red arrows highlight the structure of the true graphical model over metabolites; blue arrows correspond to those effects that are not mediated by a latent pathway.

the introductory paragraph (*i.e.* inevitable need for untestable assumptions, possible high-dimensionality).

Having given a rough description of the hurdles we must overcome, we can think about the strengths of datasets that comprise genetic and phenotypic data and make our research questions revolve around these. Namely:

- More and more, datasets contain hundreds, if not thousands, of variables:

Research Question 1: What does the large number of variables being observed enable?

In other words, what can be done with hundreds of variables that cannot be achieved when modelling the distribution of two variables? Intuitively, we imagine that this will allow us to identify patterns that would otherwise remain hidden.

- The direction of effect of some variables is known:

Research Question 2: How can we use our knowledge about the direction of some edges to increase the identifiability of our models?

Here, we want to reduce the search space by integrating our knowledge of the problem.

An interesting byproduct of Research Question 1 pertains to the form taken by the assumptions that have to be made. As we will show with an example in the next section, observing a larger set of variables sometimes makes it possible to formulate assumptions that are easier to comprehend and verify than the ones required when only a few variables are modelled.

We just sketched a model that encodes the key features of the class of datasets we are primarily interested in. We also identified the need for untestable assumptions as

one of the major challenges of graphical modelling in the presence of latent variables. We will now turn to some of the approaches published in the literature in order to see how these challenges have been tackled. We will also exemplify some of the concepts and claims that were made earlier, *e.g.* what do these untestable assumptions actually look like in practice?; how do they change when the number of observed variables increases?

1.1 Previous work

The model described in the previous section (Figure 1.2) is rather generic and too imprecise to be useful in practice since more restrictions are required, for example about the structure of the graph and the effect sizes. Here, we will review some published methods that are relevant to the problem at hand, either because they have been developed specifically to tackle that kind of problem, or because they are approaches that can be used in a wide variety of tasks and they happen to be useful in this context.

We start by recalling a few important definitions and concepts relevant to graphical models. We then give a brief description of the method of instrumental variables and explain how it relates to genetics. This framework is interesting because recent developments in the field have been making headway on Research Questions 1 and 2. This will therefore constitute a good example of a method that is capable of using many variables wisely.

In the second half, we consider a completely different family of methods that aim to estimate inverse covariance matrices using penalised maximum likelihood estimators. As will be made clearer below, inverse covariance estimation is related to Directed Acyclic Graph (DAG) estimation (at least in the Gaussian case) and enjoys very attractive computational properties.

1.1.1 Graphical models: key concepts and definitions

A graphical model is a model defined with respect to a graph whose nodes correspond to random variables and whose edges encode conditional independence statements among variables (Lauritzen 1996, Koller & Friedman 2009). These conditional independencies are of the form “ X_1 is independent of X_2 conditional on the set S ” (noted $X_1 \perp_{\mathcal{G}} X_2 | S$) and can be read off the graph using the *d-separation criterion* – whose definition is not relevant here (see Lauritzen (1996) for more details). The subscript $\perp_{\mathcal{G}}$ is here to remind us that such conditional independence statements are statements about $\mathcal{G}$ (through *d*–separation), and $\mathcal{G}$ alone. In order to relate a DAG $\mathcal{G}$ to the distributions it represents, the *Markov condition* is necessary. Under this condition, one shows that for any DAG $\mathcal{G}$ and any probability distribution P , any *d*–separation relationship that is true in $\mathcal{G}$ corresponds to a true independence relationship in P . In other words, if $X_1 \perp_{\mathcal{G}} X_2 | S$ then $X_1 \perp_P X_2 | S$ for almost all P that factorise according to $\mathcal{G}$. Here, $X_1 \perp_P X_2 | S$ is to be understood as $P(X_1 | X_2, S) = P(X_1 | S)$.

The reverse process – learning about the structure of a graph from a distribution – requires the *faithfulness* assumption. If a distribution P is *faithful* to a graph $\mathcal{G}$ and $X_1 \perp_P X_2 | S$ then $X_1 \perp_{\mathcal{G}} X_2 | S$. Thus, given n random vectors drawn from P and assuming faithfulness, one could imagine testing whether $X_1 \perp_P X_2 | S$ for all possible sets S . This would yield a set of conditional independence statements and it would then be possible to enumerate all the DAGs that are consistent with these independencies. The PC algorithm described in Spirtes et al. (2001) is an efficient and principled implementation of this idea. When some variables are hidden, the equivalent algorithm is the Fast Causal Inference (FCI) algorithm (Spirtes et al. 2001). Unfortunately, in many settings of interest (many variables, small sample size) it is unrealistic to expect the faithfulness assumption to hold (Uhler et al. 2013).

We now pursue our introduction to graphical models and introduce another useful

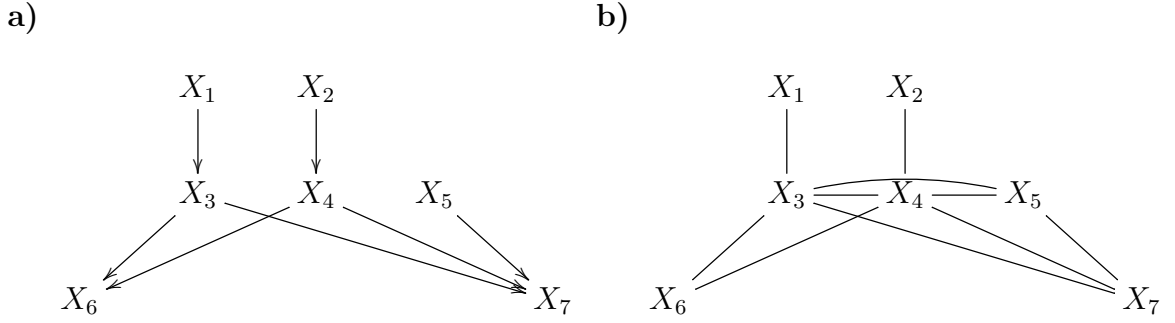


Figure 1.3: **a)** $\mathcal{G}$, a DAG. **b)** $Mor(\mathcal{G})$, the moralised counterpart of $\mathcal{G}$.

piece of notation. Let $I(\mathcal{G})$ denote the set of independencies obtained from a DAG $\mathcal{G}$ by using the d-separation criterion. One says that a DAG $\mathcal{G}$ is an *I-map* for the distribution P if all the independencies in $I(\mathcal{G})$ are satisfied by P (note that the converse is not true). For example, if $X_1 \perp\!\!\!\perp_{\mathcal{G}} X_2 | X_3$ and $\mathcal{G}$ is an I-map for P , then $P(X_1 | X_2, X_3) = P(X_1 | X_3)$. Furthermore, $\mathcal{G}$ is a *minimal* I-map for P if a) it is an I-map for P ; b) $\mathcal{G}' \subset \mathcal{G}$ implies that $\mathcal{G}'$ is not an I-map for P .

Using these definitions, we can introduce the undirected equivalent of a graphical model defined by a DAG: its *moralised* graph. To each DAG $\mathcal{G}$, we assign an undirected graph $Mor(\mathcal{G})$ obtained by adding an edge between any two nodes that are parents of the same node. In addition, all directed edges are replaced by undirected edges (see Figure 1.3 for an example). This is the correct way of defining the undirected equivalent of a DAG because it can be shown that $Mor(\mathcal{G})$ is a minimal I-map for $\mathcal{G}$. This tells us two things: i) some information might be lost in the process as we have $I(Mor(\mathcal{G})) \subseteq I(\mathcal{G})$ and not $I(Mor(\mathcal{G})) = I(\mathcal{G})$; ii) this is the best we can possibly achieve because removing any edge from $Mor(\mathcal{G})$ would induce independencies that are not in $I(\mathcal{G})$.

We conclude this brief introduction by recalling that the sets of distributions that can be described by DAGs and undirected graphical models are not identical, although their intersection is not empty.

We now consider the special case of Gaussian Graphical Models (GGMs), which are undirected graphical models. A well-known property of GGMs (also known as Gaussian Markov Random Fields) is the connection existing between the inverse covariance matrix Σ^{-1} (also known as *precision* or *concentration* matrix) and the structure of the (undirected) graphical model (V, E) : there is an edge between V_i and V_j if and only if $(\Sigma^{-1})_{ij}$ is non-zero (Lauritzen 1996). This property, coupled with the attractive computational properties of GGMs, explains the attention given to inverse covariance matrix estimation (see Section 1.1.3 for more on this topic):

- if one is interested in learning a Gaussian DAG, estimating the corresponding precision matrix gives a handle on the moralised graph. The moralised graph being a minimal I-map for the underlying DAG, it is the “closest” undirected graph to the true DAG and can be used as a starting point for other methods.
- if the true model is not a DAG, an undirected graphical model is still useful to describe the conditional independencies between observed variables.

1.1.2 Mendelian randomisation: genetic variants as instrumental variables

Instrumental variables (IV) have been used in fields such as econometrics and genetic epidemiology in order to quantify causal relationships between pairs of observed variables. When using instrumental variables one seeks to answer a causal question of the form “would an intervention on X change the distribution of Y ”. The notion of intervention is therefore central to causality and, following Pearl (2000), the new notation $\mathbb{P}(Y|do(X = x))$ will be used in order to describe an intervention that forces the variable X to take the value x . This is to be contrasted with the usual conditional distribution $\mathbb{P}(Y|X = x)$.

Different statistical frameworks to study causality have been developed over the

years and IVs have been defined in various contexts using

1. Linear Structural Models, notably in econometrics (Heckman & Robb 1985, Pearl 2000);
2. Non-parametric structural equation models (Pearl 2000);
3. Causal Directed Acyclic Graphs (Pearl 2000, Spirtes et al. 2001, Dawid 2002);
4. Counterfactuals (Angrist et al. 1996).

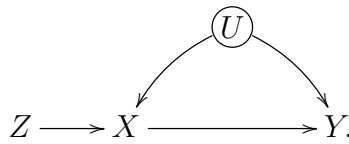
Any given statement, or definition, will usually have an equivalent in a different framework, expressed with similar, but different, concepts (Hernán & Robins 2006). Here, we define instrumental variables using a graphical model (with a causal interpretation) and, since it is enough for our purposes, we limit ourselves to the linear case.

In many cases, the quantity of interest is the *Average Causal Effect* (ACE), which is defined to be the difference in expectation between two settings of X :

$$ACE(x_1, x_2) := \mathbb{E}[Y|do(X = x_2)] - \mathbb{E}[Y|do(X = x_1)].$$

1.1.2.1 Definition

Assume that we are interested in estimating the causal effect of X on Y , and that a third random variable Z is observed. Let U be the unobserved random variable that summarises the effect of the variables that might confound the study. Then Z is called a *valid instrument* if the distribution $\mathbb{P}(Z, X, Y, U)$ factors according to the DAG



This condition can also be expressed in terms of conditional independence statements using the d –separation criterion as follows:

1. Z is independent of U : $Z \perp\!\!\!\perp U$.
2. Conditionally on X and U , Z is independent of Y : $Z \perp\!\!\!\perp Y|(X, U)$.

These are sometimes referred to as the *exclusion restriction*.

Here, because of linearity, the *average causal effect* (ACE) of X on Y is simply the coefficient one would obtain by performing a linear regression of Y on X , with U as covariate (which is impossible because U is unobserved). Under the IV setup, and assuming linearity, Z can be used to obtain an unbiased estimate of the ACE. Indeed, from Z one builds a proxy for X , $\hat{X}$ say, that is independent of U . A regression of Y on $\hat{X}$ then yields the desired quantity (Angrist & Imbens 1995).

In the context of *Mendelian randomisation*, the instrument Z is a genetic variant. Mendelian randomisation relies on the key observation that, due to genetic recombination, the genetic material of an offspring is a random sample of its ancestors. Each parent has a set of pairs of chromosomes (one from each grandparent). During meiosis, the genetic material of the grandparents is shuffled, and passed on to the offspring. Thus, both random mating and recombination play their part in the randomisation process. An example of an ideal Mendelian randomisation study is given in Chen et al. (2008) where the authors investigate the effect of alcohol intake (X) on blood pressure (Y) and use a genetic variant (Z) located in a gene regulating alcohol metabolism. Individuals with one or two copies of that variant tend to drink significantly less than individuals who do not have that allele, which is why it can be used as a proxy for alcohol consumption. This makes it possible to control for the effect of unobserved confounders such as smoking and possibly other unknown hidden variables (Chen et al. 2008). In this instance, the study concluded that alcohol is causal for an increase in blood pressure.

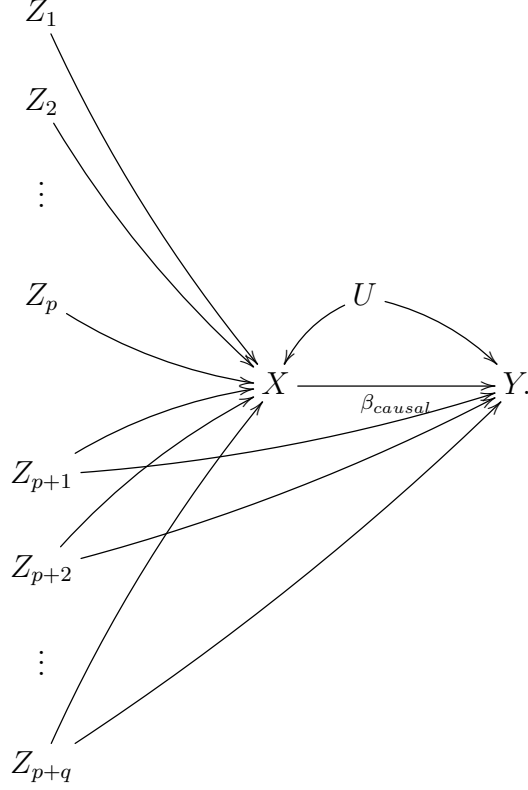
While there exist many accounts of successful Mendelian randomisation studies, it is clear that conditional independencies 1 and 2 are difficult to verify in practice since they involve U , an unobserved quantity which we know nothing about. In order to justify the use of a given genetic variant as an instrumental variable, one must therefore have an in depth *a priori* knowledge of the problem at hand, as was the case in the example of the previous paragraph.

Going back to the original problem introduced with the help of Figure 1.2, we see that, should a valid instrument be available for each of the observed phenotypes, it would be possible to use this approach in order to query the presence or absence of an edge between any pair of phenotypes. Instruments could therefore be used to reconstruct the true causal DAG, provided one carefully accounts for conditional independencies between phenotypes and assumes some form of faithfulness. Note that in this scenario, it would also be possible to recover the topological ordering of the variables.

1.1.2.2 Using multiple instruments

Finding valid instruments is challenging but it is easy, however, to make a list of *putative* instruments by scanning the genome and keeping only those variants that are significantly associated with the intermediate phenotype of interest, X . This results in a mixture of valid and invalid instruments, but no knowledge about which ones are valid and invalid is available. Let us assume that we are given a set of $p + q$ instruments, p of which are valid, q of which are invalid. Then, the true causal graph

would be:



Surprisingly, recent work has shown that if $p > q$, then the p valid instruments can be identified without prior knowledge about which instruments are the valid ones (Kang et al. 2016). Under this assumption, it also follows that a consistent estimator of β_{causal} (which is the average causal effect) is available.

It is interesting to see what multiple instruments enable. First of all, the condition that more than 50% of the instruments must be valid is much weaker than the one would usually assume, namely that all instruments be valid. Moreover, this assumption is easier to comprehend for the practitioner. It does not require an in-depth understanding of the mechanism through which each of the instruments acts on X .

Thinking back about Research Questions 1 and 2 (how can we use the fact that many variables are measured, how to best incorporate prior knowledge), this result by Kang et al. (2016) is particularly relevant. It shows how assumptions can get weakened when more variables are measured.

1.1.2.3 Limitations

As was already flagged earlier, the main limitation of IVs relates to the exclusion restriction. While carefully justifying the validity of a few instruments is possible, it quickly becomes intractable as their number increases. Thanks to the recent advances pointed out in the previous paragraph, it is sometimes possible to relax some of the IV assumptions. This method can be very useful in practice but it still cannot cope with the complications created by latent pathways: for some traits, none of the variants are valid instruments. Moreover, no matter how many instruments are being used, IV methods are still limited to *pairs* of phenotypes.

The other main drawback of instrumental variables comes from the fact that they rely on hypothesis testing. This is a limitation they share with more general algorithms such as the *Fast Causal Inference* (FCI) algorithm mentioned earlier. When analysing a dataset with a large number of variables – and *a fortiori* a high-dimensional dataset – thousands of tests are required in order to establish the structure of the graph. In that setting, it is unreasonable to expect the faithfulness assumption to hold (Uhler et al. 2013). This implies that, due to the large number of tests being performed, some errors will be made and they will quickly propagate to the entire graph.

In conclusion, the method of instrumental variables is a valuable tool when it comes to investigating predefined questions about a particular dataset but is not well suited to the exploration of a large dataset. It is preferable to use some other method in order to screen for interesting hypotheses and then to investigate these hypotheses further using IVs.

1.1.3 Inverse covariance estimation

In the previous section, we described a method which is capable of estimating causal quantities but is not well-suited to the exploration of a dataset with many variables.

Other methods which rely on the unrealistic faithfulness assumption (*e.g.* FCI) suffer from the same problem. Even when this assumption holds, only a subset of the edges can be oriented, so that only a *completed partially directed acyclic graph* (CPDAG) can be estimated. In summary, there are three problems that cannot be overcome: 1) the faithfulness assumption rarely holds; 2) due to Markov equivalence, some edges cannot be oriented; 3) the search space of CPDAGs is too large to be exhaustively explored. The latter issue could be addressed by using greedy algorithms such as the *Greedy Equivalence Search* (GES) algorithm of Chickering & Boutilier (2002), but it cannot account for latent variables and, most importantly, its computational cost does not scale well with the number of variables.

Here, we consider another family of approaches which do not address these problems directly but rather attempt to learn the structure of an undirected graph which is related to the original DAG. Such methods estimate the inverse covariance matrix of the joint distribution and, unlike the methods mentioned above, model all variables jointly. They are computationally efficient, are able to cope with high-dimensional datasets and do not rely on the faithfulness assumption. The task of inferring an inverse covariance matrix is very relevant to the problem at hand because, in the Gaussian case, it is a stepping-stone towards DAG estimation. In particular, an estimate of the inverse covariance matrix can be used in conjunction with some of the methods described in the previous paragraph.

1.1.3.1 Problem statement

Throughout, we assume that we are given n independent, identically distributed, realisations of a Gaussian random vector $Y \in \mathbb{R}^{m+p+h}$. Y is indexed by disjoint subsets of $\{1, \dots, m+p+h\}$, denoted Z, X, H and with respective cardinality m, p and h . They correspond to the variables we wish to condition on, the variables we wish to model and the hidden variables. We write Y_Z (resp. Y_X and Y_H) for the

subvector of Y indexed by Z (resp. X and H).

Σ^* denotes the true covariance matrix over Z, X, H , and we define $K^* = \Sigma^{*-1}$, the true precision matrix. For any given matrix, such as K^* , the matrices $K_Z^*, K_{ZX}^*, \dots$ represent the $m \times m, m \times p, \dots$ submatrices extracted by considering only a subset of the rows/columns, as indicated by the subscripts. Furthermore, K_O^* represents the precision matrix indexed by all observed variables : $O = Z \cup X$. Finally, we write Σ_O^n for the sample covariance matrix over the observed variables. Following the notations introduced above, Σ_Z^n is the covariance matrix over Z , while Σ_X^n is the covariance matrix over X . The superscript $\cdot^n$ is a reminder that such quantities are sample covariances formed from n realisations of Y .

1.1.3.2 Modelling all variables jointly: the graphical lasso

As already pointed out, the relationship between precision matrix and graph structure makes it possible to learn a sparse graph by controlling the number of non-zero entries of the estimate. In practice, model selection in the context of GGMs is often performed using an ℓ_1 -regularised maximum likelihood estimator (MLE) commonly known as the graphical lasso (Yuan & Lin 2007, Friedman et al. 2008). An estimate $\hat{S}$ of K^* is obtained as a solution of

$$\hat{S} = \arg \min_{S \succ 0} -\log \det S + Tr(S\Sigma_O^n) + \lambda_n \|S\|_1,$$

where $\lambda_n > 0$. The subscript n is a reminder that the optimal value of the tuning parameter varies with the sample size. Note that while the formulation given here is the most common, Yuan & Lin (2007) penalise only the off-diagonal elements. This can, in some circumstances, be more numerically stable.

The ℓ_1 -norm is the convex envelope of the ℓ_0 unit ball and is therefore a natural convex relaxation to learn sparse matrices. Building on the success of the graphical

lasso, estimators of the form “log-likelihood” + “non-Euclidian convex penalty” have received considerable interest (Chandrasekaran, Recht, Parrilo & Willsky 2012). A relevant example is the use of the nuclear norm (*i.e.* the sum of the singular values) as a convex relaxation for learning low-rank models (Bach 2008). Such regularised MLEs have attractive computational properties, for example the graphical lasso can be fitted to 1,000 variables in at most a minute (Friedman et al. 2008). The ℓ_1 and nuclear norm regularised MLEs also offer strong theoretical guarantees (Bach 2008, Ravikumar et al. 2011). For example Ravikumar et al. (2011) showed that, under well-defined circumstances, the graphical lasso is capable of correctly recovering a sparse graphical model.

Taking this problem as a starting point, two elementary operations on multinormal distributions can be performed: conditioning and marginalising. The approaches we describe now arise as an answer to the challenges raised by each of these operations.

1.1.3.3 Conditioning : Sparse Gaussian Conditional Markov Random Fields.

Sohn & Kim (2012) consider the problem of learning a GGM (over X , say) conditional on a subset of the observed variables (Z , say). It is assumed that there are no hidden variables so that $H = \emptyset$. In this context, it is well-known that

$$Y_X|Y_Z \sim \mathcal{N}(-K_X^{*-1}K_{ZX}^{*T}Y_Z, K_X^{*-1}).$$

Consequently, Sohn & Kim (2012) suggest the following estimates $(\hat{S}_{ZX}, \hat{S}_X)$ for (K_{ZX}^*, K_X^*) :

$$(\hat{S}_{ZX}, \hat{S}_X) = \arg \min_{S_X \succ 0, S_{ZX}} \ell(S_{ZX}, S_X; \Sigma_O^n) + \lambda_n(\|S_X\|_1 + \|S_{ZX}\|_1),$$

where

$$\ell(S_{ZX}, S_X; \Sigma_O^n) = -\log \det S_X + \text{Tr}(\Sigma_X^n S_X + 2S_{ZX}^T \Sigma_{ZX}^n + S_X^{-1} S_{ZX}^T \Sigma_{ZX}^n S_{ZX}),$$

and $\lambda_n > 0$. The entries of both S_X and S_{ZX} are being shrunk in order to jointly learn a pair of sparse matrices describing the direct effects of Z on X and the graph over X . Wytock & Kolter (2013) studied the theoretical properties of this estimator and derived a set of sufficient conditions for the correct recovery of K_X^* and K_{ZX}^* . Through simulations and a number of applications, Sohn & Kim (2012), Wytock & Kolter (2013), Zhang & Kim (2014) showed that, in some settings, this approach constitutes a significant improvement over the graphical lasso in both model selection and prediction tasks.

1.1.3.4 Marginalising: Low-Rank Plus Sparse Decomposition.

We now turn to the problem of learning a graphical model in the presence of hidden variables. We assume that only the variables indexed by X and Z are observed so that the marginal precision matrix is given by the Schur complement of H in K^* . Therefore,

$$Y_O \sim \mathcal{N}\left(0, (K_O^* - K_{OH}^* K_H^{*-1} K_{OH}^{*T})^{-1}\right).$$

Here K_O^* would be our target: it encodes the structure of the graphical model over O , while $K_{OH}^* K_H^{*-1} K_{OH}^{*T}$ encodes the marginal effect of all the hidden variables on O . Since the marginal precision matrix is the sum of two matrices ($S - L$, say), the problem is fundamentally misspecified. However, following the seminal work of Candès et al. (2011) and Chandrasekaran et al. (2009), Chandrasekaran, Parrilo & Willsky (2012) showed that it is sometimes possible to correctly decompose $S - L$ into its summands and recover the structure of the graph encoded by S . Loosely speaking, this is the case if K_O^* is sparse and there are relatively few hidden variables with an

effect spread over most of the nodes in O . As a result, Chandrasekaran, Parrilo & Willsky (2012) suggest a regularised maximum likelihood estimator which penalises the ℓ_1 -norm of S and the nuclear norm of L as follows:

$$(\hat{S}, \hat{L}) = \arg \min_{S-L \succ 0, L \succeq 0} \ell(S - L; \Sigma_O^n) + \lambda_n(\gamma \|S\|_1 + \|L\|_*),$$

where $\ell(K; \Sigma_O^n) = \text{Tr}(K \Sigma_O^n) - \log \det K$ and $\lambda_n, \gamma > 0$. With the help of the ℓ_1 and nuclear penalties, the precision matrix is therefore decomposed into the sum of a sparse and a low-rank matrix. Among other useful results, Chandrasekaran, Parrilo & Willsky (2012) showed that this estimator is, under suitable conditions, sparsistent and “ranksistent” : the sign patterns of both the entries of S and the spectrum of L can be recovered exactly. Modelling latent variables leads to two related, but distinct, problems:

- *identifiability* : when does the problem admit a *unique* solution? Notice that, unlike the breakdown caused by the high-dimensional regime, this kind of non-identifiability is more fundamental and remains no matter how large the number of samples.
- *consistency* : provided there exists a unique solution, when can the estimator correctly recover that solution?

In summary, in the approaches that we have described here, *conditioning* is modelled by using the relevant likelihood, while *marginalising* is handled by decomposing the marginal precision matrix into the sum of two components.

1.1.3.5 Limitations

In the introduction, we identified three main challenges (high-dimensionality, confounding and requirement to model some variables conditional on the others) and two research questions (how to best use many variables, how to use our knowledge

about the direction of some edges). The approaches described above made significant advances on these challenges and questions. For example, the low-rank plus sparse approach accounts for confounding by detecting the footprint of latent variables across all observed variables. In that respect, this method makes a clever use of the large number of variables being observed. Unfortunately, the limitation on the number of hidden variables – for which the rank of the low-rank piece is a proxy – can be quite restrictive. On the other hand, the graphical lasso and the sparse conditional GGM approach of Wytock & Kolter (2013) require that all the relevant variables be observed so that none of the approaches published in the literature is capable of coping with the full problem.

1.2 Contributions

Chandrasekaran, Parrilo & Willsky (2012) and Sohn & Kim (2012) made significant progress on the challenges mentioned in the introductory paragraph but none of these methods can cope with the full problem. Here, I combine both approaches in order to learn a sparse Gaussian conditional random field in the presence of latent variables. Inputs (variables in Z) are allowed to act on X in both a sparse and a low-rank fashion while the inverse covariance matrix over X is estimated conditional on Z and the latent variables. In this setting, latent factors can either mediate the effects of Z on X or simply act as confounders on X . By doing so I generalise the work done in Chandrasekaran, Parrilo & Willsky (2012), Sohn & Kim (2012), but also provide an alternative answer to some of the problems recently considered in the multivariate regression literature, *e.g.* reduced-rank regression with inverse covariance estimation (Chen & Huang 2016). As will be shown later, this approach makes it possible to correctly recover graphs that are typically denser and with more hidden variables than the ones that can be handled by other methods.

My contributions are the following:

- Building on the framework of Chandrasekaran, Parrilo & Willsky (2012), I study the theoretical properties of a novel estimator (defined in chapter 2). In particular, I establish that it is well-behaved in the high-dimensional setting.
- I describe algorithms that can be used when the number of variables is in the thousands. More specifically, I show how the objective function (and the one of Wytock & Kolter (2013)) can be recast as a Semi-Definite Programming (SDP) problem.
- Through simulations, I illustrate the conditions required for identifiability and show that this approach is consistent in a wider range of settings than the approaches introduced earlier.
- I devise a metric which uses an independent source of information in order to assess the biological relevance of our estimates.
- I apply the method (and competing approaches) to two real-life datasets. In the first application, I use the proposed approach to model the levels of 39 metabolic traits conditional on hundreds of genetic variants, in two independent cohorts. I find my results to be better replicated across cohorts than those obtained with other methods. In the second application, I look at a high-dimensional gene expression dataset. I find that my method is capable of retrieving as many biologically relevant gene-gene interactions as other methods while retrieving less irrelevant interactions.

1.3 Outline

The thesis is organised as follows.

The next chapter is dedicated to the definition of the estimator we will study during the rest of this thesis. We seize this opportunity to recall our problem statement.

In Chapter 3 we investigate its theoretical properties: conditions required for identifiability, convergence rate, etc... This chapter focuses only on key properties and results, the proofs are deferred to Appendix B.

In Chapter 4 we discuss the practical aspects of our work. We explain how the model is fitted, how the tuning parameters are chosen and how we score our estimates in order to assess their biological relevance and allow fair comparisons between our approach and other methods.

Chapter 5 illustrates the behaviour of the suggested estimator through simulations. In particular, we investigate how the number of latent variables impacts our ability to recover the nominal parameters.

Chapter 6 and 7 are application chapters. We study two very different datasets: one well-powered metabolites dataset and one high-dimensional gene expression dataset.

Finally, in the discussion chapter we reflect on the results of our simulations and applications and suggest new directions of research.

Chapter 2

Problem Statement and Suggested Estimator

In the introductory chapter, we described three approaches that aim to estimate an inverse covariance matrix from multivariate normal data. In light of that work, we suggest decomposing the parameters of a Gaussian conditional Markov random field into the sum of a low-rank and a sparse matrix.

2.1 Problem statement

Throughout, we assume that we are given n independent, identically distributed, realisations of a random vector $Y \in \mathbb{R}^{m+p+h}$. Y is indexed by disjoint subsets of $\{1, \dots, m+p+h\}$, denoted Z, X, H and with respective cardinality m, p and h . They correspond to the variables we wish to condition on, the variables we wish to model and the hidden variables. We write Y_Z (resp. Y_X and Y_H) for the subvector of Y indexed by Z (resp. X and H). We do not make any assumptions about the distribution of Y_Z . However, the random vector $\begin{pmatrix} Y_X \\ Y_H \end{pmatrix}$ is assumed to be a Gaussian random vector, whose conditional mean vector is a function of Y_Z .

Σ^* denotes the true covariance matrix over Z, X, H , and we define $K^* = \Sigma^{*-1}$, the true precision matrix. For any given matrix, such as K^* , the matrices $K_Z^*, K_{ZX}^*, \dots$ represent the $m \times m, m \times p, \dots$ submatrices extracted by considering only a subset of the rows/columns, as indicated by the subscripts. Furthermore, K_O^* represents the precision matrix indexed by all observed variables : $O = Z \cup X$. Finally, we write Σ_O^n for the sample covariance matrix over the observed variables. Following the notations introduced above, Σ_Z^n is the covariance matrix over Z , while Σ_X^n is the covariance matrix over X . The superscript $\cdot^n$ is a reminder that such quantities are sample covariances formed from n realisations of Y .

Throughout, we make use of a number of standard notations (*e.g.* $\| - \|_1$ for the ℓ_1 -norm). A list of notations and their definitions can be found in Appendix A.

2.2 Suggested estimator

As explained in the introduction, we are interested in modelling X conditional on Z , under the assumption that the variables indexed by H are unobserved. Three relevant estimators from the literature were introduced:

1. the graphical lasso (GLASSO) (Friedman et al. 2008): this estimator models all observed variables jointly without making any distinction between X and Z . Furthermore, it assumes that there no latent variables.
2. the sparse Gaussian conditional Markov random field (SCGGM) (Wytock & Kolter 2013): it improves on GLASSO by modelling X conditional on Z . Latent variables are not modelled.
3. the low-rank plus sparse approach (Chandrasekaran, Parrilo & Willsky 2012): this estimator models $X \cup Z$ in the presence of latent variables.

Thus, SCGGM and LRPS offer only *partial* solutions to the problem at hand.

To better address our problem and design a novel estimator, we suggest combining the key features of these two methods. In order to understand what this means in practice, we recall here the form taken by SCGGM and LRPS.

Conditioning: The sparse Gaussian conditional Markov random field (SCGGM) assumes a model of the form

$$Y_X|Y_Z \sim \mathcal{N}(-K_X^{*-1}K_{ZX}^{*T}Y_Z, K_X^{*-1}),$$

and estimates the parameters (K_{ZX}^*, K_X^*) by minimising

$$(\hat{S}_{ZX}, \hat{S}_X) = \arg \min_{S_X, S_{ZX}} \ell(S_{ZX}, S_X; \Sigma_O^n) + \lambda_n(\|S_X\|_1 + \|S_{ZX}\|_1),$$

subject to

$$S_X \in \mathbb{R}^{p \times p}, S_X \succ 0, S_{ZX} \in \mathbb{R}^{m \times p},$$

and where

$$\ell(S_{ZX}, S_X; \Sigma_O^n) = -\log \det S_X + \text{Tr}(\Sigma_X^n S_X + 2S_{ZX}^T \Sigma_{ZX}^n + S_X^{-1} S_{ZX}^T \Sigma_Z^n S_{ZX}).$$

Marginalising: On the other hand, the low-rank plus sparse approach (LRPS) assumes that

$$Y_O \sim \mathcal{N}\left(0, (K_O^* - K_{OH}^* K_H^{*-1} K_{OH}^{*T})^{-1}\right),$$

and computes estimates $(\hat{S}, \hat{L})$ of $(K_O^*, K_{OH}^* K_H^{*-1} K_{OH}^{*T})$ by optimising

$$(\hat{S}, \hat{L}) = \arg \min_{S, L} \ell(S - L; \Sigma_O^n) + \lambda_n(\gamma \|S\|_1 + \|L\|_*),$$

subject to

$$S, L \in \mathbb{R}^{(p+m) \times (p+m)}, S - L \succ 0, L \succeq 0,$$

where $\ell(K; \Sigma_O^n) = \text{Tr}(K \Sigma_O^n) - \log \det K$ and $\lambda_n, \gamma > 0$.

Building on the work mentioned above, we suggest the following estimator, which decomposes the parameters S_{ZX}, S_X of SCGGM into sums of low-rank and sparse matrices:

$$\begin{aligned}
(\hat{S}_X, \hat{L}_X, \hat{S}_{ZX}, \hat{L}_{ZX}) = & \\
& \arg \min_{S_X, L_X, S_{ZX}, L_{ZX}} \ell(S_{ZX} - L_{ZX}, S_X - L_X; \Sigma_O^n) + \lambda_n(\gamma \|S\|_1 + \|L\|_*) \\
& \text{s.t } S_X - L_X \succ 0, L_X \succeq 0 \text{ and } S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix},
\end{aligned}$$

(2.1)

where

$$\ell(K_{ZX}, K_X; \Sigma_O^n) = -\log \det K_X + \text{Tr}(\Sigma_X^n K_X + 2K_{ZX}^T \Sigma_{ZX}^n + K_{ZX} K_X^{-1} K_{ZX}^T \Sigma_Z^n),$$

and with

$$S_X, L_X \in \mathbb{R}^{p \times p}, S_{ZX}, L_{ZX} \in \mathbb{R}^{m \times p}.$$

Solving (2.1) amounts to solving a function which is *jointly convex* in its parameters over a *convex constraint set* (proofs are in the appendix): an operation which can in general be performed in polynomial time. As mentioned earlier, this likelihood is structured around two parameters, K_{ZX} and K_X , accounting respectively for the direct (*i.e.* conditional on all variables) effects of Z on X and the structure of the graph over X . However, because we penalise the rank of L , the effect of all latent variables is modelled *jointly* and a single set of latent factors is learned. No distinc-

tion is being made between the variables that mediate the action of Z and the ones that act as confounders on X . On the other hand, the parameters S_X and S_{ZX} retain their interpretability.

The rest of this thesis is dedicated to the study of the theoretical and computational properties of (2.1). We also assess its performances on two real-life datasets.

Chapter 3

Theoretical Results

In the previous chapter, we introduced a novel estimator and showed that it arises naturally from the study of independent identically distributed (i.i.d.) realisations of a Gaussian vector, after conditioning on part of the variables and assuming that some variables are not observed. We now turn to the theoretical properties of our approach.

For the time being, we assume that each sample is generated according to the model

$$Y_X|Y_Z \sim \mathcal{N}\left(-(S_X^* - L_X^*)^{-1}(S_{ZX}^* - L_{ZX}^*)^T Y_Z, (S_X^* - L_X^*)^{-1}\right), \quad (3.1)$$

and ask under what circumstances our estimator correctly recovers the parameters S^*, L^* (as built by stacking S_X^*, S_{ZX}^* and L_X^*, L_{ZX}^*) with overwhelming probability. Note that the distribution of Y_Z does not play any part in our analysis. In particular, Y_Z can very well be a discrete random variable.

We analyse this problem in the framework of Chandrasekaran, Parrilo & Willsky (2012) and therefore our proofs often mirror theirs. However, because of the form taken by the likelihood and because we do not limit ourselves to square matrices, the analysis is more involved. Our goal being to provide the reader with an intuition for the range of applicability of our method, we choose to defer most technical details

to the appendices and give only a few key results. As the reader will soon realise, we use a significant number of notations, definitions and constants. In order to make this chapter and our proofs easier to follow, we make available a list of notations and their definitions in Appendix A.

As mentioned earlier, modelling latent variables by decomposing the parameters into a sum of two matrices raises *identifiability* issues: given samples drawn from (3.1), when is it possible to exactly decompose the sum $S + L$ (where S, L are defined as before) into its summands? This is a problem which has been tackled in great generality in Chandrasekaran, Parrilo & Willsky (2012) and their results directly apply to the present situation. To make the present work self-contained, we start by recalling a few notations and key definitions that will be used throughout the rest of this document. Then, an overview of the conditions required for identifiability is given and we look at our main result: the *consistency* of (2.1).

3.1 Key definitions and assumptions

Until now, we have repeatedly mentioned the fact that a “low-rank plus sparse decomposition” was possible if S is sparse and L is low-rank. However, it is clear that imposing conditions on the sparsity of S and the rank of L is not sufficient. For example, consider the matrix with a single entry: it is at the same time sparse and low-rank and there is therefore no unique way of decomposing it into the sum of a low-rank and a sparse matrix. Chandrasekaran et al. (2009) introduce the notion of *rank-sparsity incoherence* and define quantities that make it possible to express clearly the conditions under which such a problem is well-posed. More details about the motivation behind these definitions are given in Chandrasekaran, Parrilo & Willsky (2012).

3.1.1 Key definitions

We write $\mathcal{S}(k)$ for the algebraic variety of $(m+p) \times p$ matrices with at most k non-zero entries. Any $(m+p) \times p$ matrix M with support of cardinality k is a smooth point of $\mathcal{S}(k)$ and we denote by $\Omega(M)$ the tangent space to $\mathcal{S}(k)$ at M (this is the set of matrices in $\mathcal{S}(k)$ whose support is included in the support of M).

Similarly, we define $\mathcal{L}(r)$, the variety of $(m+p) \times p$ matrices with rank at most r . Any matrix M of rank r is a smooth point of $\mathcal{L}(r)$ and we write $T(M)$ for the tangent space to $\mathcal{L}(r)$ at M (this is the set of matrices in $\mathcal{L}(r)$ whose row space is identical to M 's or whose column space is identical to M 's).

Given two linear spaces T_1, T_2 of identical dimensions, write

$$\rho(T_1, T_2) \triangleq \max_{\|N\|_2 \leq 1} \|(\mathcal{P}_{T_1} - \mathcal{P}_{T_2})(N)\|_2$$

for a measure of the “angle” between T_1 and T_2 . Here $\mathcal{P}_{T_1}$ is the orthogonal projector onto T_1 while $\|N\|_2$ is the spectral norm of N (*i.e.* its largest singular value). $\rho(T_1, T_2)$ is used to describe the set of tangent spaces that are close to the nominal $T(L^*)$.

Chandrasekaran et al. (2009) quantify the spread of the effect of the latent variables on the observed variables by defining

$$\xi(T(M)) \triangleq \max_{N \in T(M), \|N\|_2 \leq 1} \|N\|_\infty$$

for any matrix M . Remark that $\xi(T(M))$ controls all the elements of the tangent space at M . Thus, if $\xi(T(M))$ is small, none of the matrices having the same row-space (resp. column-space) as M can have a large ℓ_∞ -norm, meaning that the row-space (resp. column-space) of M cannot be closely aligned with any of the coordinate axes. Therefore, a small $\xi(T(M))$ guarantees that no single latent variable will have a strong effect on only a small set of the observed variables (precisely because it is

the observed variables that define the coordinate axes).

Similarly, define

$$\mu(\Omega(M)) \triangleq \max_{N \in \Omega(M), \|N\|_\infty \leq 1} \|N\|_2,$$

to quantify the diffusivity of the spectrum of M is. One shows that matrices with a small number of non-zero entries per row/column (and thus sparse) have a small $\mu(M)$. In particular, $\mu(\Omega(M)) \leq \text{degree}(M)$.

3.1.2 Conditions on the Fisher Information Matrix (FIM)

In the context of the graphical lasso, the *mutual incoherence* or *irrepresentability* condition is a well-known requirement for identifiability (Ravikumar et al. 2011). Just like the irrepresentability condition imposes restrictions on the Hessian of the likelihood, the conditions given in Chandrasekaran, Parrilo & Willsky (2012) involve the Fisher Information Matrix (FIM) $\mathcal{I}_{\Sigma_Z^n}^*$ evaluated at the true parameters $S^* - L^*$. The subscript is a reminder that the entire analysis is performed conditional on Σ_Z^n . To avoid cluttered notations, we write $\Omega = \Omega(S^*)$ and $T = T(L^*)$ and denote by $\mathcal{P}_\Omega, \mathcal{P}_T$ the orthogonal projections onto these linear subspaces.

We can now define the quantities that control the behaviour of the FIM when restricted to Ω and T . Set

$$\begin{aligned} \alpha_\Omega &\triangleq \min_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_\Omega \mathcal{I}_{\Sigma_Z^n}^* \mathcal{P}_\Omega(M)\|_\infty; \\ \delta_\Omega &\triangleq \max_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_{\Omega^\perp} \mathcal{I}_{\Sigma_Z^n}^* \mathcal{P}_\Omega(M)\|_\infty, \\ \beta_\Omega &\triangleq \max_{M \in \Omega, \|M\|_2 = 1} \|\mathcal{I}_{\Sigma_Z^n}^*(M)\|_2. \end{aligned}$$

Likewise let,

$$\begin{aligned}\alpha_T &\triangleq \min_{\rho(T, T') < \xi(T)/2} \min_{M \in T', \|M\|_2=1} \|\mathcal{P}_{T'} \mathcal{I}_{\Sigma_Z^n}^* \mathcal{P}_{T'}(M)\|_2; \\ \delta_T &\triangleq \max_{\rho(T, T') < \xi(T)/2} \max_{M \in T', \|M\|_2=1} \|\mathcal{P}_{T'^\perp} \mathcal{I}_{\Sigma_Z^n}^* \mathcal{P}_{T'}(M)\|_2, \\ \beta_T &\triangleq \max_{\rho(T, T') < \xi(T)/2} \max_{M \in T', \|M\|_\infty=1} \|\mathcal{I}_{\Sigma_Z^n}^*(M)\|_\infty.\end{aligned}$$

Finally, we define

$$\alpha \triangleq \min(\alpha_\Omega, \alpha_T) \quad \beta \triangleq \max(\beta_\Omega, \beta_T) \quad \delta \triangleq \max(\delta_\Omega, \delta_T).$$

Then the following assumption is a generalisation of the irrepresentability condition, and from now on we assume that it holds.

Assumption 1. (*Generalised Irrepresentability Condition, Chandrasekaran, Parrilo & Willsky (2012)*)

There exists a $\nu \in (0, \frac{1}{2}]$ such that

$$\frac{\delta}{\alpha} \leq 1 - 2\nu.$$

A direct consequence of Assumption (1) is that $\alpha > 0$, so that the FIM is injective on Ω and on all the spaces that are close to T (including T itself). For that reason, Assumption (1) subsumes the standard *restricted convexity assumption* which is also necessary in the context of sparse conditional Gaussian Markov random fields and regularised regression in order to guarantee that the likelihood – when restricted to the true support – is strictly convex, and therefore that there exists a unique optimum (Wytock & Kolter 2013, Wainwright 2009).

Another consequence of Assumption 1 is Proposition 1 given in Appendix B.2. This proposition is applied multiple times throughout the consistency proof. It holds only if the following conditions are met. This shows the role played by μ and ξ in

identifiability.

Assumption 2. (*Assumptions for Proposition 1*)

$$\mu(\Omega)\xi(T) \leq \frac{1}{6} \left(\frac{\nu\alpha}{\beta(2-\nu)} \right)^2$$

and γ is chosen in the range

$$\gamma \in \left[\frac{3\xi(T)\beta(2-\nu)}{\nu\alpha}, \frac{\nu\alpha}{2\mu(\Omega)\beta(2-\nu)} \right].$$

Unsurprisingly, a sparse S^* (small $\mu(\Omega)$) and a diffuse effect of the latent variables on the observed nodes (small $\xi(T)$) increase the chances that Assumption (2) will hold.

It is legitimate to ask whether there exist interesting classes of matrices that satisfy Assumption 2. This problem is addressed in details in Chandrasekaran et al. (2009). In particular, we have the following theorem (Chandrasekaran et al. (2009), Corollary 4).

Theorem 1. (*Chandrasekaran et al. (2009), Lemma 1*)

Let A be a $p \times p$ matrix which is such that its support is chosen uniformly at random from the collection of all support sets of size k_1 . Let B be a rank- k_2 $p \times p$ matrix, with SVD given by $B = UDV^T$ and for which U and V are chosen uniformly from the set of partial isometries of size $k_2 \times p$.

Then, $\mu(\Omega)\xi(T) \leq \frac{1}{6}$ with high-probability, provided

$$k_1 \lesssim \frac{p^{1.5}}{\log p \sqrt{\max(k_2, \log p)}}.$$

It is interesting to see that k_1 is allowed to increase at a rate which is faster than p . Thus, for a given k_2 , the number of non-zero entries of the sparse component can grow super-linearly in p . We will give more details about the number of samples

required to estimate such models after the statement of the consistency theorem.

3.2 Consistency

We can now present our main result and state the consistency of Estimator (2.1) (see Appendix B for the proof).

We define the following quantities : $D = \max(1, \frac{\nu\alpha}{3\beta(2-\nu)})$, $\psi_Z = \|\Sigma_Z\|_2$, $\psi_X^* = \|K_X^{*-1}\|_2$ and $\phi_{ZX}^* = \|K_{ZX}^*\|_2$, $\psi = \frac{3}{2}\psi_X^* \sqrt{\left(1 + 2\frac{\psi_Z}{\psi_X^*} \left(1 + \frac{9}{4}\psi_X^* \phi_{ZX}^*\right)^2\right)}$. We also set

$$\begin{aligned} C_1 &= \frac{48}{\alpha} + \frac{1}{\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right)}, \\ C_2 &= \left(1 + \frac{24(2-\nu)}{\nu}\right) C_1^2 D \psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right), \\ C_3 &= C_1 + \frac{3\alpha C_1^2 (2-\nu)}{4(3-\nu)}, C_4 = \max\{C_2, C_3\}, C_5 = \frac{C_1 \nu \alpha}{\beta(2-\nu)}, \\ C_6 &= \frac{\alpha \nu}{32(3-\nu)D} \min\left(\frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha \nu}{384D(3-\nu)\psi_X^* \psi^2 (1 + \frac{\alpha}{6\beta})^2}\right). \end{aligned}$$

Finally, set:

$$\delta_n = \sqrt{\frac{256\psi_X^{*2} p M}{n}},$$

with $M = \max\left(1, \frac{\psi_Z}{4\psi_X^*} (1 + \sqrt{\frac{m}{p}})^2\right)$. And, let

$$\lambda_n = \frac{6D(2-\nu)\delta_n}{\xi(T)\nu}.$$

Then we prove the following theorem in the appendix:

Theorem 2. (*Algebraic Consistency*)

Suppose that Assumptions 1 and 2 hold and that we are given n samples drawn according to (3.1). Further assume that the following hold:

$$1. \ n \geq \frac{pM}{\xi(T)^4} \max\left(2, \frac{256\psi_X^{*2}}{C_6^2}\right).$$

2. Let the minimum non-zero singular value of L^* be such that

$$\sigma \geq \frac{C_4 \lambda_n}{\xi(T)^2}.$$

3. Let the minimum magnitude nonzero entry θ of S^* be such that

$$\theta \geq \frac{C_5 \lambda_n}{\mu(\Omega)}.$$

Then, with probability greater than $1 - 2 \min \left(\exp(-pM), \exp\left(-\frac{4\psi_X^*}{\psi_Z} pM\right) \right)$, we have that

$$1. \text{ sign}(\hat{S}) = \text{sign}(S^*) \text{ and } \text{rank}(\hat{L}) = \text{rank}(L^*)$$

2.

$$\max \left(\frac{1}{\gamma} \left\| \hat{S} - S^* \right\|_{\infty}, \left\| \hat{L} - L^* \right\|_2 \right) \leq \frac{32(3-\nu)}{3\alpha(2-\nu)} \lambda_n.$$

In particular, due to the form taken by δ_n ,

$$\max \left(\frac{1}{\gamma} \left\| \hat{S} - S^* \right\|_{\infty}, \left\| \hat{L} - L^* \right\|_2 \right) \leq C \frac{1}{\xi(T)} \sqrt{\frac{pM}{n}},$$

$$\text{with } C = \frac{1024(3-\nu)D\psi_X^*}{\nu\alpha}.$$

Seen at a high-level, this result is analogous to the one obtained for the low-rank plus sparse decomposition of Chandrasekaran, Parrilo & Willsky (2012) (*e.g.* scaling regime, dependency on ξ, μ), but our analysis also reveals important features of the problem that are specific to this setting.

Among others, there are some important consequences of this theorem:

1. This result holds even if m and p grow as a function of n . More specifically, the required number of samples n grows linearly with pM , while the probability of algebraic consistency increases exponentially in pM .

2. Both the identifiability and consistency results depend on the structure of the graph and the number of hidden variables only through their dependence on $\mu(\Omega)$ and $\xi(T)$.
3. Conditions c) and d) are fairly standard for this kind of estimator but can be quite restrictive in practice. Through their dependency on λ_n , they vary with $\frac{1}{\sqrt{n}}$. By analogy with the well-known β_{min} condition required for the consistency of the lasso (in regression analysis), we will sometimes refer to them as the σ_{min} and β_{min} conditions.

We remark that because we model all latent variables jointly (with L) our analysis does not really distinguish between S_X, L_X, L_{ZX} and S_{ZX} . Specifically, our proof makes a recurrent use of the identity

$$\max(\|A\|_2, \|B\|_2) \leq \left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_2 \leq \sqrt{2} \max(\|A\|_2, \|B\|_2),$$

hence the form taken by δ_n . Whether it may be possible to decompose the contribution of these terms is not yet clear.

Prior to stating Theorem 2 we showed that natural ensembles of matrices satisfy Assumption 2. One also shows the following results (Chandrasekaran, Parrilo & Willsky 2012, Section 4.2).

Throughout, we assume that the low-rank matrix L^* is almost maximally incoherent: $\xi(T(L^*)) = \mathcal{O}\left(\sqrt{\frac{h}{p+m}}\right)$ (the random matrix set mentioned earlier satisfies this condition). Then, we have:

- if $\deg(S^*) = \mathcal{O}(1)$ and $h = \mathcal{O}(p+m)$ then the model can be estimated with $n = \mathcal{O}(p+m)$ samples.
- if $\deg(S^*) = \mathcal{O}(\log(p+m)^q)$ and $h = \mathcal{O}\left(\frac{p+m}{\log(p+m)^{2q}}\right)$ then the model can be consistently estimated from $n = \mathcal{O}((p+m) \text{polylog}(p+m))$.

It is thus possible to increase the degree of S^* and the number of variables as a function of n . For example, provided the S^* has fixed degree, the number of latent variables can grow *linearly* with n .

Chapter 4

Practical Considerations

Having investigated the theoretical properties of the suggested estimator, we now turn to the challenges raised by applying our method (and other similar regularised maximum likelihood estimators) to real data. Namely, a) finding the minimizer of the objective function given the observed covariance matrix, b) setting the tuning parameters to appropriate values and c) assessing the quality of estimated structures so as to allow for fair and relevant comparisons between methods.

For convenience – and to understand why task a) and b) are challenging – let us first recall that the estimates $\hat{S}_X, \hat{S}_{ZX}, \hat{L}_X, \hat{L}_{ZX}$ are defined in terms of the solution to

$$\begin{aligned}
 (\hat{S}_X, \hat{L}_X, \hat{S}_{ZX}, \hat{L}_{ZX}) = \\
 \arg \min_{S_X, L_X, S_{ZX}, L_{ZX}} \ell(S_{ZX} - L_{ZX}, S_X - L_X; \Sigma_O^n) + \lambda_n(\gamma \|S\|_1 + \|L\|_*) \\
 \text{s.t } S_X - L_X \succ 0, L_X \succeq 0 \text{ and } S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}, \quad (4.1)
 \end{aligned}$$

where Σ_O^n and ℓ are defined as before.

As already stated earlier, (4.1) is jointly convex in its parameters, thus providing

guarantees about the recovery of its global minimizer. However, while convex, the penalty terms are not smooth, which prevents the use of standard optimisation algorithms (*e.g.* the Newton–Raphson method and its derivatives). Due to the popularity of lasso-type approaches, similar optimisation problems have been extensively studied and many satisfactory algorithms can be found in the literature. However, because it is not straightforward to apply such algorithms to the present setting, we give details on the issues that are specific to the form taken by (4.1). Since they offer different performances in different settings, we investigate two approaches: proximal gradient algorithms and semidefinite programming.

Task b) concerns the choice of the tuning parameters λ_n and γ . The first parameter – λ_n – controls the overall shrinkage and is often encountered in regularised likelihood estimators. The literature on choosing λ_n is therefore extensive and there exist many competing approaches (the most popular one being cross-validation). However, in biostatistics it is often required to report estimates subject to some form of error control. For that reason, we use a method known as *complementary pairs stability selection* in our applications. In the second part of this chapter, we give details about this method, how we choose λ_n and select the edges that make up the final estimate. The other tuning parameter, γ , is less common and, as we have experienced in our own simulations, cross-validation is not helpful to select its value. Here, we suggest a strategy which monitors how sensitive the estimates are to the value of γ (after stability selection). We assess the performances of this scheme on simulated data. The protocol established in this chapter is the one we follow in application chapter 6.

Finally, problem c) relates to the challenge of comparing the performances of different approaches, in the context of applications to real data. To that end, we suggest a score for graphical structures which relies on external sources of information (*e.g.* ontologies). Such external data is used to compute an enrichment statistic reflecting whether a given structure tends to connect “related” nodes (in a sense that

will be made clear later on) more often than would be expected in a random graph with a similar topology. This performance metric will be used in both application chapters (Ch. 6, 7). The last section of the present chapter is dedicated to its definition.

4.1 Fitting

Here we regard (4.1) as an optimisation problem and suggest a strategy to fit the model when m and p are in the thousands. Due to the popularity of convex relaxation approaches involving non-smooth penalties, this class of problem has been extensively studied in the literature. A popular family of algorithms – proximal algorithms (PA) – relies on one’s ability to efficiently compute the proximal operator of the non-smooth component of the likelihood (Parikh & Boyd 2014). As will be made clear below, the proximal operator required to fit (4.1) cannot be computed in closed form but it is nonetheless possible to efficiently obtain a good approximation to it. Another option consists in recasting (4.1) into a semidefinite program (SDP) (Vandenberghe & Boyd 1996). In this context, this is an attractive alternative to PAs due to the availability of solvers that are capable of handling large-scale datasets (Wang et al. 2010). Solvers such as the ones described in (Wang et al. 2010, Tütüncü et al. 2003) make it possible to fit objective functions comprising log-determinant terms and complex combinations of linear, quadratic and semidefinite constraints. Although both PAs and SDPs ought to yield the same solution (and they very often do), we noticed that PAs are preferable when the inputs Z act on X in a sparse fashion, with few or no mediating latent variables. In our experience, this observation also holds when fitting the low-rank plus sparse estimator of Chandrasekaran, Parrilo & Willsky (2012) using Split-Bregman iterations (Ye et al. 2011) vs. SDPT3 (Tütüncü et al. 2003) (which itself appears to be sometimes more accurate than LogdetPPA (Wang et al. 2010)).

We note in passing that other approaches such as the alternating direction of multipliers (ADMM) and the (block) coordinate descent algorithm have been developed in order to solve similar problems (Boyd 2011, Osher & Li 2009). Unfortunately, the ADMM is not well-suited to the problem at hand, since one of the steps cannot be efficiently computed. We have not investigated the properties of coordinate descent for this particular application, but unlike ADMM it has never been suggested to solve the Low-Rank plus Sparse problem.

In this section, we omit the details that are common to all non-smooth convex optimisation problems and focus on the challenging parts of either method that are specific to the suggested estimator:

- computing the proximal operator of our non-smooth penalty under the constraints imposed by (4.1);
- recasting (4.1) to make it suitable for semidefinite programming techniques.

Both implementations are made available online ¹.

4.1.1 Using a proximal gradient algorithm

We focus on a difficulty which is specific to (4.1): the lack of closed-form solution for a certain proximal operator.

Seen at a high level, (4.1) is a well-studied problem of the form

$$\text{minimize } f(x) + g(x) \tag{4.2}$$

where $f : \mathbb{R}^k \rightarrow \mathbb{R}$, $g : \mathbb{R}^k \rightarrow \bar{\mathbb{R}}$ are closed proper convex functions and f is differentiable.

¹<https://github.com/benjaminfrot/lasso-type-estimators>

In this context, we have

$$f : (K_{ZX}, K_X) \mapsto -\log \det K_X + \text{Tr} (\Sigma_X^n S_X + 2K_{ZX}^T \Sigma_{ZX}^n + K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX}),$$

and

$$g : (K_{ZX}, K_X) \mapsto \lambda(\gamma \|S\|_1 + \|L\|_*) + \mathcal{I}_{\mathcal{S}^{\succ 0}}(S - L) + \mathcal{I}_{\mathcal{S}^{\succeq 0}}(L),$$

where $S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}$ and $L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}$. Here we defined $\mathcal{I}_{\mathcal{S}^{\succ 0}}$ as the indicator function of the convex set

$$\mathcal{S}^{\succ 0} \triangleq \left\{ \begin{pmatrix} K_X \\ K_{ZX} \end{pmatrix} \mid K_X \in \mathbb{R}^{p \times p}, K_{ZX} \in \mathbb{R}^{m \times p}, K_X \succ 0 \right\}.$$

$\mathcal{I}_{\mathcal{S}^{\succeq 0}}$ is defined similarly for positive semidefinite matrices.

The non-differentiability of g prevents the resort to standard optimisation methods, which is why (4.2) is typically solved using some form of proximal algorithm (Parikh & Boyd 2014). The simplest instance of such an algorithm is a proximal gradient method with iteration

$$x^{k+1} := \text{prox}_{\eta^k g}(x^k - \eta^k \nabla f(x^k)), \quad (4.3)$$

where η^k is some step size and where $\text{prox}_{\eta g}$ is the *proximal operator* of ηg :

$$\text{prox}_{\eta g}(v) \triangleq \arg \min_x \eta g(x) + \frac{1}{2} \|x - v\|_2^2.$$

This is a well-known algorithm which has been extended and improved in many ways (*e.g.* Nesterov's method which was used to solve a similar problem in Zhang & Kim (2014), Nesterov (2005)). However, no matter which flavour of the algorithm is being used, computing the proximal operator fast and accurately is critical. In particular,

Schmidt et al. (2011) show how convergence is affected when an error is present in the calculation of the proximal operator and provide guidelines about the number of inner iterations required to achieve the optimal convergence rate.

This is relevant to our setting because one shows that applying iteration (4.3) to solve (4.1) requires the computation of a proximal operator of the form

$$\text{prox}_{\mathcal{I}_{\mathcal{S}^{\geq 0}}(\cdot) + \lambda \|\cdot\|_*}(X) = \arg \min_L \mathcal{I}_{\mathcal{S}^{\geq 0}}(L) + \lambda \|L\|_* + \frac{1}{2} \|L - X\|_F^2, \quad (4.4)$$

which is the proximal operator of the sum of two functions : one is the indicator function of a convex set, while the other corresponds to the regularisation imposed on the nuclear norm of L . Here $\|L - X\|_F$ denotes the Frobenius norm of the matrix $L - X$. Considered independently, the proximal operators of $\mathcal{I}_{\mathcal{S}^{\geq 0}}(L)$ and $\lambda \|L\|_*$ admit well-known closed form solutions that rely on the thresholding of the eigenvalues of L_X and the singular values of L , respectively (Parikh & Boyd 2014). The proximal operator of their sum, however, cannot be computed in closed form. We must resort to the “Dykstra-like” proximal algorithm of Combettes & Pesquet (2009) which generates a sequence that converges to the solution of (4.4). See Algorithm 1 for more details.

Initialisation: Set $X_0 = X$, $P_0 = 0$, $Q_0 = 0$;

for $k = 1, \dots, K$ **do**

$Y_k = \text{prox}_{\mathcal{I}_{\mathcal{S}^{\geq 0}}}(X_k + P_k);$ $P_{k+1} = X_k + P_k - Q_k;$ $X_{k+1} = \text{prox}_{\lambda \ \cdot\ _*}(Y_k + Q_k);$ $Q_{k+1} = Y_k + Q_k - X_{k+1}$
--

end

Result: X_k .

Algorithm 1: “Dykstra-like” algorithm of (Combettes & Pesquet 2009). The sequence X_k converges to the solution of problem (4.4). In practice, $\frac{\|X_{k+1} - X_k\|_F}{\|X_k\|_F} \leq \epsilon$ is used as a stopping criterion. We choose a fixed ϵ of 10^{-5} , but an adaptive scheme might result in faster convergence rates (Schmidt et al. 2011). Typically, each iteration has time complexity $O(\max(p, m)^3)$.

As a result, each iteration of (4.3) requires the computation of the gradient of f

(with time complexity $O(p^3)$), the proximal operator of the ℓ_1 -norm which is a simple thresholding of the entries and a few iterations of Algorithm 1 – typically less than 10.

In practice, we use a Nesterov’s second method – which is a form of accelerated proximal gradient algorithm – that we complement with the adaptive restart method introduced in O’Donoghue & Candès (2013), Nesterov (2005). In spite of the difficulties raised by the form of (4.4) this approach provably converges to the optimum of (4.1) (Schmidt et al. 2011).

4.1.2 Recasting the objective function as a Semidefinite Program (SDP)

The solvers made available in the MATLAB[®] packages SDPT3 and Logdet-PPA are capable of solving problems of the form (Tütüncü et al. 2003, Wang et al. 2010):

$$\arg \min_{X_1, X_2, \dots} Tr(X_1 C_1^T) + Tr(X_2 C_2^T) + \dots + a_1 \log \det(X_1) \quad (4.5)$$

subject to a number of linear, quadratic and positive semidefinite constraints². Our goal is then to recast (4.1) as a problem of the same form as (4.5). This involves introducing additional constrained variables, as shown below.

First, we focus on the form taken by the likelihood of a Gaussian conditional random field (GCRF). Estimating a GCRF amounts to solving

$$\arg \min_{K_X, K_{ZX}} -\log \det K_X + Tr(\Sigma_X^n K_X + 2K_{ZX}^T \Sigma_{ZX}^n + K_{ZX} K_X^{-1} K_{ZX}^T \Sigma_Z^n) \quad (4.6)$$

with $K_X \succ 0$. This differs from (4.5) by the presence of the term $K_{ZX} K_X^{-1} K_{ZX}^T$ which involves the inverse of the parameter K_X . To address this problem, we introduce a

²See references for a formulation of the problem in its full generality.

new variable W and solve the equivalent problem

$$\arg \min_{K_X, K_{ZX}} = -\log \det K_X + \text{Tr}(K \Sigma_O^n)$$

such that

$$K = \begin{pmatrix} W & K_{ZX} \\ K_{XZ} & K_X \end{pmatrix} \succeq 0, \text{ and } K_X \succ 0.$$

This formulation is equivalent to (4.6) because Σ_Z^n is a covariance matrix and therefore positive-semidefinite. As a result, the constraint $K \succeq 0$ is equivalent to $W = K_{ZX} K_X^{-1} K_{ZX}^T$ ³.

Having addressed this issue, there remains the problem of recasting the penalty $\lambda_n(\gamma \|S\|_1 + \|L\|_*)$ as a combination of linear constraints and trace terms. One shows that (Chandrasekaran et al. 2009):

$$\arg \min_{S, L} \gamma \|S\|_1 + \|L\|_*$$

is equivalent to

$$\arg \min_{S, L, F, H_1, H_2} \lambda_n \left(\gamma \mathbf{1}^T F \mathbf{1} + \frac{1}{2} (\text{Tr}(H_1) + \text{Tr}(H_2)) \right)$$

subject to

$$\begin{pmatrix} H_1 & L \\ L^T & H_2 \end{pmatrix} \succeq 0, \quad -F_{ij} \leq S_{ij} \leq F_{ij}, \quad \forall i, j;$$

where the latter constraint is to be understood elementwise.

³We thank Prof. Defeng Sun for his help on this problem.

Putting everything together, we obtain the following SDP formulation of (4.1):

$$\begin{aligned}
& \arg \min_{S_X, L_X, S_{ZX}, L_{ZX}, W, F, H_1, H_2} Tr(K \Sigma_O^n) - \log \det S_X + \lambda_n \left(\gamma \mathbf{1}^T F \mathbf{1} + \frac{1}{2} (Tr(H_1) + Tr(H_2)) \right) \\
& \text{subject to } K \succeq 0, S_X \succ 0, L_X \succeq 0, \begin{pmatrix} H_1 & L \\ L^T & H_2 \end{pmatrix} \succeq 0, -F_{ij} \leq S_{ij} \leq F_{ij}, \forall i, j; \\
& \text{where } K = \begin{pmatrix} W & S_{ZX} - L_{ZX} \\ S_{ZX}^T - L_{ZX}^T & S_X - L_X \end{pmatrix}, S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}.
\end{aligned} \tag{4.7}$$

Problem (4.7) can then easily be implemented using *e.g.* YALMIP and solved using LogdetPPA or SDPT3 (Löfberg 2004, Wang et al. 2010, Tütüncü et al. 2003).

4.2 Selecting the tuning parameters

Because regularised maximum likelihood estimators such as the suggested approach and the lasso simultaneously perform inference and variable selection, they can be used in a wide variety of tasks. In some cases, one might be interested in maximising predictive power and not give much attention to which variables are entering the model. In another context, it might be the opposite: the estimated effect sizes are discarded and only the variables being selected are of importance. Naturally, these two problems are related but the task at hand often decides on how the tuning parameters are chosen. For example, cross-validation is well-suited for prediction tasks but tends to select too many edges. For that reason, what constitutes an “appropriate” value of the tuning parameters is very much application dependent.

In biostatistics, a lot of attention is given to reproducibility: it is usually required to have some control over the number of associations that are falsely discovered. Here, we use *complementary pairs stability selection*, a very general method that

can be applied to any variable selection algorithm and which provides, under some conditions, bounds on some error rates for false discoveries (Meinshausen & Bühlmann 2010, Shah & Samworth 2013). It addresses two known shortcomings of lasso-type estimators: the high sensitivity to the tuning parameters (especially λ_n) and the lack of closed-form sampling distribution of the estimates.

In this section, we describe the procedure that we follow in chapter 6. For any given value of the tuning parameter γ , stability selection allows us to choose λ_n and return an estimate that comes with some guarantees on the number of falsely discovered edges. We then show that our approach is not very sensitive to γ and provide guidelines on how to pick its value. Through simulations, we illustrate the behaviour of the proposed procedure.

4.2.1 Stability selection

Stability selection was first introduced in Meinshausen & Bühlmann (2010) and then perfected in Shah & Samworth (2013), where *complementary pairs stability selection* was defined. It is a general technique that aims at improving the performances of variable selection algorithms by considering random subsamples of the data. We refer the reader to Meinshausen & Bühlmann (2010), Shah & Samworth (2013) for technical details on these methods and now give an overview of both stability selection and complementary pairs stability selection. In particular, we draw the attention of the reader to the fact that their results bound different quantities and hold for different assumptions.

Throughout this section, we assume that we have a variable selection algorithm that selects a subset of the variables indexed by $\{1, \dots, w\}$. For example, in the context of graphical model selection, if there are p variables, $w = \binom{p}{2}$ since this is the number of edges in the complete undirected graph over p nodes. We denote the set of selected edges for a given value of the tuning parameter λ by $S_n^\lambda \subseteq \{1, \dots, w\}$.

Remark that for any λ , S_n^λ is implicitly a function of the samples $I = \{1, \dots, n\}$. Consequently, for any $J \subseteq I$, we write $S^\lambda(J)$ for the set obtained by applying the variable selection method to the subset of the data indexed by J .

We can now define the stability selection ($\hat{S}_{SS,n}^\lambda$) and complementary pairs stability selection ($\hat{S}_{CPSS,n}^\lambda$) versions of the original selection procedure. Let $\mathcal{J} = \{J_1, \dots, J_B\}$ be a collection of B subsets of I such that $|J_i| = \lfloor \frac{n}{2} \rfloor, \forall i \in \{1, \dots, B\}$. Then, for $k \in \{1, \dots, w\}$, one defines

$$\hat{\Pi}(k)^\lambda := \frac{1}{B} \sum_{j=1}^B \mathbb{1}_{k \in \hat{S}^\lambda(J_j)}.$$

In words, $\hat{\Pi}(k)^\lambda$ describes the empirical inclusion probabilities of the w possible edges: the base method is applied to B subsets of the data and we count how often each edge has been selected. Stability selection and complementary pairs stability selection refer to two different sampling schemes for $\mathcal{J}$:

- Stability Selection: $\mathcal{J}$ is built by sampling B random subsets of I with replacement.
- Complementary Pairs Stability Selection: the B subsets are generated by sampling $B/2$ pairs of subsets such that $J_{2i} \cap J_{2i+1} = \emptyset$, for $i \in \{1, \dots, B/2\}$.

We then have the following definition for $\hat{S}_{SS,n}^\lambda$ and $\hat{S}_{CPSS,n}^\lambda$.

Definition 1. (*(Complementary Pairs) Stability Selection*)

Let $0 \leq \tau \leq 1$ be a threshold on the inclusion probabilities. Then we define

$$\hat{S}_{SS,n}^{\lambda,\tau} := \{k : \hat{\Pi}_{SS}^\lambda(k) \geq \tau\}, \quad \hat{S}_{CPSS,n}^{\lambda,\tau} := \{k : \hat{\Pi}_{CPSS}^\lambda(k) \geq \tau\}.$$

These quantities describe the edges that are included with probability at least τ , for a given value of λ . They are often referred to as the stable edges.

Intuitively, we can imagine that stable edges are more likely to be part of the true set S^* of edges, while noise variables will be included in only a few folds. Moreover, the choice of the tuning parameter λ and the choice of the threshold τ are intimately related – with τ tempering the effect of λ . The following theorem of Shah & Samworth (2013) makes it possible to choose τ so as to achieve a bound on the *number of low probability edges* being selected.

Theorem 3. (*(Shah & Samworth 2013), Theorem 1*)

For $0 \leq \theta \leq 1$, let L_θ denote the set $\{k : p_{k, \lfloor \frac{n}{2} \rfloor} \leq \theta\}$, that is the set of edges with low inclusion probability. Here $p_{k, \lfloor \frac{n}{2} \rfloor}$ is the true inclusion probability of edge k . Then, for $\tau \in (0.5, 1]$,

$$\mathbb{E}|\hat{S}_{CPSS,n}^\lambda \cap L_\theta| \leq \frac{\theta}{2\tau - 1} \mathbb{E}|\hat{S}_{\lfloor n/2 \rfloor}^\lambda \cap L_\theta|.$$

Remark that this theorem shines by its lack of assumptions and it is therefore expected to work very well on real data. Moreover, this result holds even when B is as small as 2, although a typical choice would be $B = 100$. Under stronger assumptions, it is possible to bound the number of falsely discovered edges. This is a result proved in Meinshausen & Bühlmann (2010) for the stability selection sampling scheme and derived as a special case of Theorem 3 for the CPSS sampling scheme (Shah & Samworth 2013).

Theorem 4. (*Error Control (Meinshausen & Bühlmann 2010, Shah & Samworth 2013)*)

Let $q_\lambda := \mathbb{E}|\hat{S}^\lambda(\lfloor \frac{n}{2} \rfloor)|$, be the expected number of edges selected when applying the base selection procedure to subsets of size $\lfloor \frac{n}{2} \rfloor$. Assume that

1. the base selection procedure is better than random guessing;
2. the distribution of $\{\mathbb{1}_{k \in \hat{S}^\lambda}, k \in N\}$ is exchangeable, where N is the set of noise variables (variables that are not in the true model).

Then, the expected number of selected noise variables V is bounded above as follows:

$$\mathbb{E}(V) \leq \frac{1}{2\tau - 1} \frac{q_\lambda^2}{w}.$$

A common criticism of Theorem 4 is the presence of assumption 2 which is very strong and most likely never met on real data, hence the work of Shah & Samworth (2013) to provide a more realistic framework for stability selection. There is thus no reason to believe that this result will hold on in our applications. However, deciding on an upper bound for $\mathbb{E}[V]$ is convenient because it provides guidelines on how to choose λ and makes it possible to compare methods, for a given value of that bound, irrespective of whether it actually holds. Indeed, if the assumptions of Theorem 4 do not hold, this bound has a different interpretation: using the value of θ suggested in Shah & Samworth (2013), it follows from Theorem 3 that this quantity bounds the expected number of low probability edges being selected instead of the expected number of noise variables. We will therefore make use of the bound of Theorem 4 since it always has a meaningful interpretation and allows for a fair comparison between methods.

In order to select λ , we use *pointwise stability selection* whereby Theorem 4 is applied independently for each value of $\lambda \in \Lambda$. We typically choose 10 or 20 reasonable values of λ and then set the threshold τ on the inclusion probabilities $\hat{\Pi}_{CSS}^\lambda$ so that $\mathbb{E}(V) \leq 1$. When λ is too small, the desired bound cannot be achieved with a $\tau \leq 1$. For that reason, we keep only those estimates for which there exists a $\tau \in (0.5, 1]$ such that the bound $\mathbb{E}(V) \leq 1$ can be achieved. Of all the values of λ that yield a valid estimate, the one with the most edges is kept. It is clear that the size of the set Λ in which λ takes its values has an impact on the final estimate. Meinshausen & Bühlmann (2010) show how to account for this but nonetheless choose to use pointwise stability selection in their application. As long as $|\Lambda|$ is reasonable, 10 or

20 say, the number of values being tested should not have a major impact on the final results.

We conclude this section by remarking that none of the results above offer guarantees on the *expected number of stable edges* at a given τ : Theorem 4 cannot be used to compute False Discovery Rates (FDRs) akin to the ones commonly used in genetics (using, *e.g.* the Benjamini-Hochberg procedure).

4.2.2 Sensitivity to γ

The low-rank plus sparse method of Chandrasekaran, Parrilo & Willsky (2012) and the suggested estimator both have two tuning parameters, λ_n and γ . As already mentioned, γ controls the trade-off between low-rank and sparse components and, according to the theory, there is a *range* of values of γ for which the estimator is consistent. As will be shown through simulations, this result has two practical consequences:

- since there is a range of values of γ within which both the sign pattern and the rank of the latent piece do not change, it should be possible to study how the solutions vary with γ and identify this stable region. This is, of course, an asymptotic result and it remains to be seen whether this is a useful strategy in small sample sizes.
- the estimates are not very sensitive to this tuning parameter: finding the optimal γ is not critical as long as its value lies close enough to the right region.

As it is more convenient to have a parameter that takes values in a bounded interval we rewrite the penalty

$$\lambda_n(\gamma \|S\|_1 + \|L\|_*),$$

where $\lambda_n > 0$ and $\gamma \in (0, +\infty)$, as

$$\lambda_n(\gamma \|S\|_1 + (1 - \gamma) \|L\|_*),$$

where $\lambda_n > 0$ and $0 \leq \gamma \leq 1$. This is the form in use in all our applications and simulations. From now onwards, γ will always take values in the unit interval.

Finally, we remark that there are three ways of decomposing $S^* - L^*$ into a sum of a low-rank and a sparse matrix. Consequently, there are in general three stable regions:

1. $(\hat{S}, \hat{L}) = (S^* - L^*, 0)$: the latent component is empty and everything is explained by $\hat{S}$. This is the decomposition which is observed for small values of γ . As a matter of fact, for $\gamma \rightarrow 0$ our approach is equivalent to the sparse conditional random field of Wytock & Kolter (2013).
2. $(\hat{S}, \hat{L}) = (S^*, L^*)$: γ is within the acceptable range and the decomposition has both a sparse and a latent component.
3. $(\hat{S}, \hat{L}) = (0, S^* - L^*)$: there is no sparse component, everything is explained by the latent piece. This happens when γ is close to 1.

In the general case, and in the limit of large sample sizes, one should therefore expect three stable regions delimited by two borders within which the estimates vary very quickly with γ (see Chandrasekaran et al. (2009) for a concrete example). In practice, it might happen that there are either few latent variables (or many of them) and these regions tend to merge, in which case boundaries become hard to distinguish. However, as we will show in simulations, stable intervals are often good indicators of the optimal range for γ .

4.2.3 Suggested procedure

In the previous sections, we recalled a number of theoretical results that provide hints on how to select λ_n and γ . We now put everything together and propose an algorithm that can be followed in order to choose these values in practice. The outline of the suggested approach is simple:

1. Start with two sets (Λ, Γ) of reasonable values for λ and γ . For example, for a given value of γ , $\max(\Lambda)$ can be the smallest value of λ such that $\hat{S}^\lambda = \{\}$ ($\min(\Lambda)$ can be chosen similarly). Γ could be the set $\{1/20, 2/20, \dots, 19/20\}$.
2. For each value of $\gamma \in \Gamma$, apply pointwise complementary pairs stability selection: for each value of λ , set the threshold τ on the inclusion probabilities so that $\mathbb{E}(V) \leq 1$ (if there is no $\tau \leq 1$ such that $\mathbb{E}(V) \leq 1$, discard this estimate). Of all the valid structures, keep the densest ⁴.
3. For each value of γ , an optimal graph $\mathcal{G}(\gamma)$ has now been selected. For any pair $(\gamma_1, \gamma_2) \in \Gamma \times \Gamma$, the similarity between $\mathcal{G}(\gamma_1)$ and $\mathcal{G}(\gamma_2)$ is computed by considering the Jaccard Index of their edge sets:

$$\mathcal{J}(\mathcal{G}_1, \mathcal{G}_2) := \frac{|\text{Edges}(\mathcal{G}_1) \cap \text{Edges}(\mathcal{G}_2)|}{|\text{Edges}(\mathcal{G}_1) \cup \text{Edges}(\mathcal{G}_2)|}.$$

For all pairs $(\gamma_1, \gamma_2) \in \Gamma \times \Gamma$, compute $\mathcal{J}(\mathcal{G}(\gamma_1), \mathcal{G}(\gamma_2))$. This yields a matrix of similarities taking values in $[0, 1]$.

4. Plot that matrix as a heatmap (for example) and identify a region within which all the estimates are close to each other according to $\mathcal{J}(\cdot, \cdot)$.

We follow this procedure for both the low-rank plus sparse method and our estimator.

⁴Here we chose $\mathbb{E}(V) \leq 1$, which can be very restrictive when the number of nodes is large. Another value, 10 say, could be chosen when the number of edges is expected to be in the hundreds.

4.2.4 Performance on synthetic data

In order to assess and illustrate the suggested approach, we study its behaviour on synthetic data.

4.2.4.1 Data generation

A single simulation design is used. Here, $|Z| = |X| = 30$ and we set $|H| = 5$, so that we model the distribution of 30 “outputs” conditional on 30 “inputs”, in the presence of 5 latent variables. Each of the n samples is drawn according to a model of the form

$$X_i \sim \mathcal{N} \left(-S_X^{-1} \begin{pmatrix} S_{ZX}^T Z_i \\ S_{HX}^T H_i \end{pmatrix}, S_X^{-1} \right),$$

so that sample X_i is generated conditionally on Z_i and H_i . The inputs, Z , and the latent factors, H , are all i.i.d. and drawn from a standard normal distribution.

The random 30×30 matrix S_{ZX} has an expected sparsity of 10% and its non-zero coefficients are generated uniformly at random from the interval $[-1, 1]$. Likewise, S_{HX} is set to have an expected sparsity of 80% with non-zero coefficients drawn from $\mathcal{U}(-1, 1)$. Finally, S_X has a sparsity of 2%, with non-zero entries taken from $\mathcal{U}(-1, 1)$.

This design, with a relatively small number of latent variables connected to many observed ones, is consistent with the expectation of a low-rank plus sparse decomposition. However, since S_{ZX} is fairly dense and since the effect sizes are drawn from a uniform distribution, estimating S_X remains challenging. In particular, many effect sizes are small, and there is therefore very little chance that the β_{min} and σ_{min} conditions required for consistency will hold.

This experiment is repeated 30 times for $n \in \{1000, 10000\}$, so that there are 60 datasets in total.

4.2.4.2 Results

The relevance of the estimates is assessed using *precision* and *recall*. More specifically, we use these metrics to compare the non-zero pattern of S_X to the one of our estimates so that precision and recall have the following meanings:

- precision: the fraction of the retrieved edges that are in the graph described by S_X .
- recall: the fraction of the edges of S_X that are retrieved.

In order to graphically assess the performance of a binary classification method, it is customary to plot a precision/recall curve depicting how precision changes as a function of recall. For lasso-type methods, one typically varies λ and computes these metrics at each point of the regularisation path. In this particular instance, since there is a precision/recall curve for each value of γ , a *precision/recall surface* is used. By applying our method to all n samples and plotting this surface, it is possible to visualise what the best achievable precision/recall pair is – the one we would obtain should we be able to perfectly choose λ and γ . This can then be compared to the precision/recall actually achieved when using the procedure described in section 4.2.3.

To understand the expected behaviour of our approach, we first look at the average values taken by our metrics. Subplots **a)** and **b)** of Figure 4.1 show the precision/recall surface averaged over all 30 experiments, for $n = 1000$ and $n = 10,000$. Solid black lines mark the “optimal” range for γ : the one for which our method would achieve its best recall at a precision of at least 0.95. As can be seen from subplots **a)** and **b)**, the average optimal range spans approximately one-third of the unit interval. As expected from the theory, the final estimates are not very sensitive to the value of γ .

In each experiment, we used complementary stability selection (CPSS) (with $B = 10$) in order to select one graphical structure per value of γ . Subplots **a)**, **b)**, **e)** help us understand how CPSS performs in practice. In **a)** and **b)** red markers indicate

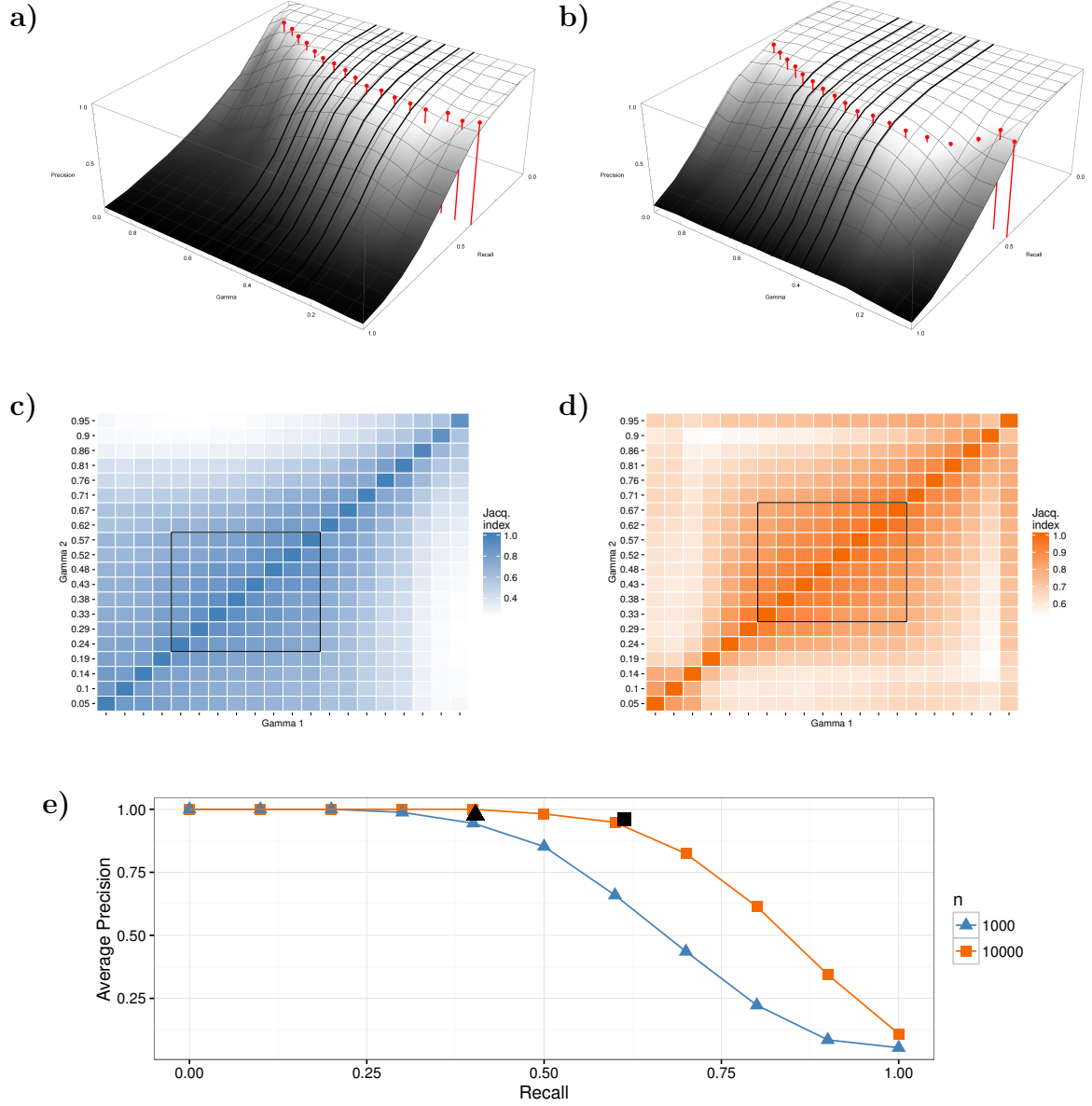


Figure 4.1: **a)** Precision/recall surface averaged over all 30 experiments, for $n = 1000$. The solid black lines mark those values of γ for which the best recall is achieved, at a precision of at least 0.95. In red, the average precision/recall obtained by applying stability selection at any given value of γ . **b)** Same as a), with $n = 10,000$. **c)** Average value of $\mathcal{J}(\mathcal{G}(\gamma_1), \mathcal{G}(\gamma_2))$, for all pairs $(\gamma_1, \gamma_2) \in \Gamma \times \Gamma$. $n = 1000$. **d)** Same as c), with $n = 10,000$. **e)** Average precision/recall curve at the optimal value of γ . In black, the precision/recall achieved using stability selection at the optimal γ .

the average precision/recall achieved using CPSS, as a function of γ . Furthermore, subplot **e)** displays the average precision/recall curve at the optimal value of γ : for each of the 30 experiments, we selected the precision/recall curve at the value of γ that yielded the highest recall at a precision of 1 and then averaged these curves. The red markers indicate the average precision/recall obtained from CPSS, at the optimal γ . There are at least two important lessons to be learned from these:

- **a), b), e)** show that, on average, one achieves a better precision and a better recall using CPSS than by applying the base method to the entire dataset. Even though the optimal value of λ is unknown, CPSS makes it possible to improve the performances of the original variable selection method. While surprising, this is consistent with the theoretical results proved in Meinshausen & Bühlmann (2010) where they show that the stability selection (SS) version of a variable selection method enjoys better consistency properties than its non-SS equivalent.
- **a), b)** show that there are in average three regimes for γ . If γ is chosen too small, a good recall is achieved but false edges enter the model (precision ≤ 1). If γ is chosen too large, one obtains conservative estimates where fewer true edges are recovered but a good precision is achieved. When γ is in the right range, both precision and recall are optimal. From the position of the red markers, we can see that the acceptable region for γ is fairly large however.

Having looked at the behaviour of CPSS, we now turn to the problem of choosing γ . Subfigures **c), d)** display the average similarity matrices obtained by computing the Jaccard index $\mathcal{J}(\mathcal{G}(\gamma_1), \mathcal{G}(\gamma_2))$ for every possible pair (γ_1, γ_2) . The solid black lines delimit the region marked by the black lines on the surface plots: these are the optimal regions for γ . It is clear that the average stable regions are spread out but their epicentre lies within the area delimited by the black rectangles.

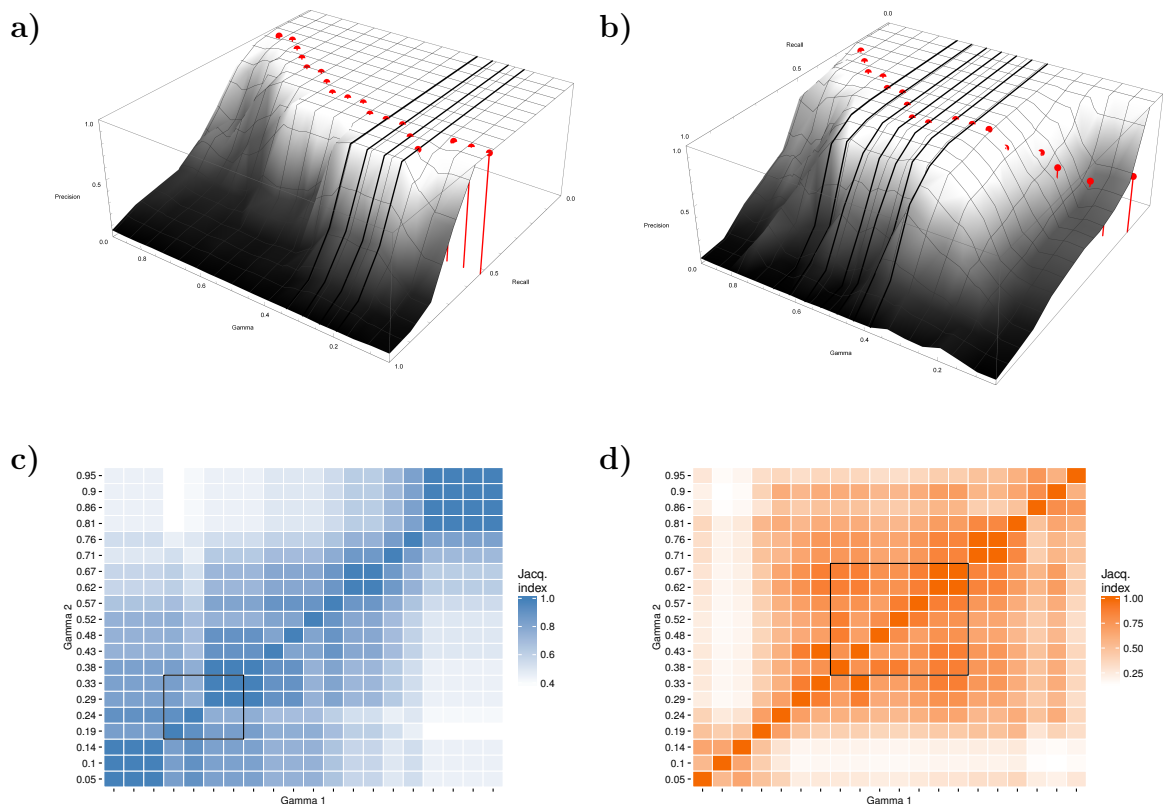


Figure 4.2: **a)** Precision/recall surface for one random experiment, for $n=1000$. The solid black lines indicate those values of γ for which the best recall is achieved, at a precision of at least 0.95. In red, the precision/recall obtained by applying CPSS, at each value of γ . **b)** Same as a), with $n = 10,000$. **c)** Value of $\mathcal{J}(\mathcal{G}(\gamma_1), \mathcal{G}(\gamma_2))$ for all pairs (γ_1, γ_2) . $n = 1000$. **d)** Same as c), with $n = 10,000$.

Figure 4.1 focuses on the results obtained on *average*: average precision/recall, average similarities, etc.... but a good average performance does not necessarily entail that there is enough information in a single dataset to correctly pick λ and γ . Figure 4.2 shows the results obtained for two random experiments, one for each sample size.

First of all, Figures 4.2 **a)**, **b)** clearly show that γ has a stronger influence than the one suggested by the average precision/recall curves: there is a substantial advantage in picking its value correctly. However, there is still a range of optimal values. As shown by the red markers, this observation still holds when using CPSS. In general, choosing γ too small results in a low precision, with many false edges being included.

A high value of γ tends to yield conservative estimates, at the cost of a poor recall. As before, Subfigures **c)**, **d)** plot the pairwise similarities for all pairs (γ_1, γ_2) . Inferring the correct region (delimited by the solid black lines) is more challenging when looking at a single dataset. **c)** provides an example of a dataset in which all three stable regions are observed. There is some overlap between the “intermediate” stable region and the valid range but, in this instance, using our procedure would not lead to choosing the optimal value of γ , although it would lead to reasonably good performances. Due to the increase in sample size, the optimal region of subfigure **d)** is much wider and therefore easier to infer. Generally speaking, a larger sample size increases the recall of the base selection procedure but also makes the method less sensitive to γ .

In conclusion, simulations show that CPSS offers good performances when it comes to choosing λ and selecting a graphical structure – for a given γ . On the other hand, while we can expect the sensitivity to γ to yield good results in average, choosing the value of that tuning parameter remains challenging. Interpreting the similarity heatmaps is not always straightforward. However, choosing γ too small really is the main concern here, as it leads to false discoveries. In that respect, our procedure is satisfactory since the stable region corresponding to the decomposition $(\hat{S}, \hat{L}) = (S^* - L^*, 0)$ is usually easily identifiable. In the worst case, one might choose too high a value of γ , which will yield conservative estimates.

4.3 Assessing the biological relevance of graphical structures

Another important step is assessing the biological relevance of the estimates using an external source of information. In bioinformatics and epidemiology, *ontologies* (*e.g.* ChEBI, the gene ontology (GO), the human phenotype ontology (HPO)) are often

available and constitute a summary of the knowledge contained in the literature on a particular topic. In this section, we define a simple metric that uses such ontologies in order to score graphs when the true structure is unknown (as is the case in real-life applications).

4.3.1 Definition

Throughout, we assume that we are interested in learning a graphical model describing a joint distribution over p variables, $\{X_1, \dots, X_p\}$ say. Since our final estimates are graphical structures (unweighted undirected graphs), we also refer to the variables X_i as *nodes* or *vertices*. We assume further that we are given a trustworthy ontology which is relevant to the problem under study. For example, the gene ontology is suitable to the study of a gene expression dataset, while ChEBI is preferable for the analysis of metabolite data. Ontologies are graphs themselves, but here they are regarded as mappings from nodes to sets of *terms*. For example, in the case of the gene ontology, each gene is assigned a set of GO IDs of the form $\{GO : 001, GO : 002, \dots\}$.

The first step – which is common to all graphs with this node set – is to match the p nodes to their equivalent in the ontology. When the variables represent genes, there is no ambiguity since each variable has usually already been given a gene symbol. Unfortunately, in some cases only a manual matching is reliable (*e.g.* when matching a variable named “DHA” to its equivalent – “Docosahexaenoic acid” – in the ChEBI ontology). At the end of this step, each vertex has been assigned a set of terms. Let $\mathcal{T} = \{T_1, T_2, \dots\}$ denote the set of all the terms T_i that annotate at least two of the p nodes.

We now assume that we are given an estimate $\mathcal{G} = (E, V)$. We annotate the edges E of $\mathcal{G}$ with the intersection of their endpoints’ annotations. For example, if node A is annotated by $\{T_1, T_2, T_3\}$ and node B is annotated by $\{T_3, T_4, T_5\}$, an edge between A and B will be assigned the set $\{T_3\}$.

Having annotated both the vertices and edges of $\mathcal{G}$, we define $\mathcal{G}(T_i) = (E(T_i), V(T_i))$ to be the subgraph of $\mathcal{G}$ restricted to those nodes and edges that have T_i in their annotation set.

For a fixed term T_i , the enrichment S_i is computed by considering the ratio of the density of $\mathcal{G}(T_i)$ to the density of $\mathcal{G}$. That is:

$$S_i := \frac{|E(T_i)|}{\binom{|V(T_i)|}{2}} / \frac{|E|}{\binom{|V|}{2}}.$$

Finally, the enrichment statistic of $\mathcal{G}$ is obtained by averaging S_i over $\mathcal{T}$:

$$\mathcal{S} := \frac{1}{|\mathcal{T}|} \sum_i S_i. \quad (4.8)$$

In practice, it is also possible to compute an empirical (one-tailed) p-value associated to this metric by permuting the annotations of $\mathcal{G}$. To be more precise, we permute the entire annotation sets (*e.g.* in a graph with two nodes A, B annotated with $\{T_1, T_2\}$ and $\{T_2, T_3\}$ respectively, a permutation $A \leftrightarrow B$ would yield the same graph but with annotations $\{T_2, T_3\}$ for A and $\{T_1, T_2\}$ for B). Thus, the null distribution is defined by a graph with the same topology as $\mathcal{G}$: vertex degrees, etc... are preserved. With such a null distribution, the complete graph always has a p-value of 1 (and a score $\mathcal{S}$ of 1).

We remark that this approach is close to the ontology analyses frequently encountered in computational biology, *e.g.* the Network Ontology Analysis (NOA) method of Wang et al. (2011).

4.3.2 Interpretation

Given the definition of $\mathcal{S}$, it is now clear why we refer to this quantity as an *enrichment statistic*: it reflects the propensity of a graph to draw connections between related nodes more often than we would expect in a random graph over the same node set

and the same number of edges. A random graph has an expected statistic of 1. Thus, for a graph to be high-scoring, it is not enough to simply connect vertices with similar annotations. For example, if a term annotates most of the nodes of $\mathcal{G}$, it will not be rare to encounter edges with that term. On the other hand, if a term is specific to only a small subset of V , there will be a substantial benefit in linking these nodes.

This concept is illustrated with a simple example in Figure 4.3. In this example, there are six nodes annotated by subsets of $\{T_1, \dots, T_5\}$. Here, T_1 and T_5 are common terms since they annotate 4 nodes each, while T_2 is very rare. Figure 4.3 **a)** gives an example of a high-scoring graph: it tends to have edges between vertices that have rare terms. Subfigure **b)** contains an example of a graph with the same number of edges but with a very low score (and a high p-value, since we use a one-tailed p-value).

Figure 4.3 **c)** and **d)** show the null distributions for the graphs depicted in **a)** and **b)**. Their empirical means is 1, as is expected from the definition of $\mathcal{S}$. A problem with this null distribution is that it is bimodal. When comparing graphs it might therefore be preferable to compare their p-values since a random graph can easily be assigned a score in the range $[0.8, 1.3]$. In our experience, however, graph selection methods perform better than random so that estimates are well above that range: comparing their enrichment statistics is rarely misleading.

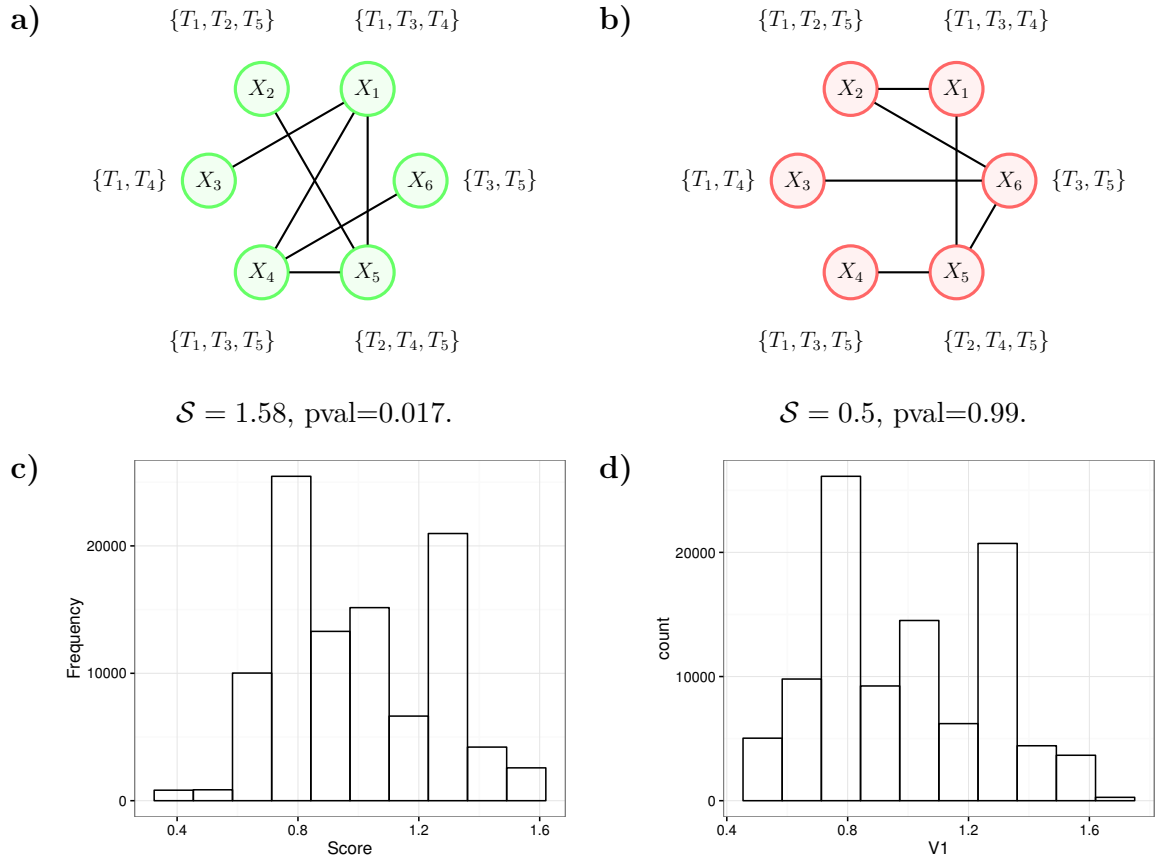


Figure 4.3: **a)** An example of a high-scoring graph. **b)** An example of a low-scoring graph. **c)** The (empirical) null distribution for the graph depicted in a). Mean of the null=1.00179, number of permutations=100,000. **d)** The null distribution for b). Mean of the null=1.000413.

Chapter 5

Simulations

We study the properties of the proposed model on synthetic data and compare its performances to the three other methods introduced earlier: the graphical lasso (Friedman et al. 2008), the sparse conditional Gaussian random Markov field (Zhang & Kim 2014, Sohn & Kim 2012, Wytock & Kolter 2013) and the low-rank plus sparse decomposition (Chandrasekaran, Parrilo & Willsky 2012). Each of these methods is subject to identifiability issues which can be caused either by a high-dimensional breakdown or because multiple models can describe the data equivalently well (*i.e.* are Markov equivalent). Here, our focus is on the latter issue: we draw a large number of samples from a range of models and measure their ability to recover the structure of the underlying graph. Even though we defined algebraic consistency in terms of our ability to recover *both* the sign pattern of S and the rank of L , more attention is given to the recovery of S .

5.1 Graphical structures

For all simulations, each observation is generated according to a model of the form

$$Y_X|Y_{ZH} \sim \mathcal{N}(-S_X^{*-1}S_{ZX}^{*T}Y_{ZH}, S_X^{*-1}).$$

To account for the sparse/low-rank behaviour of the inputs and the hidden variables we generate Y_Z and Y_H according to various models that we define below. These simulations are motivated by a desire to discriminate between the identifiability /consistency properties of the different methods. To that end, we compare methods across a minimal set of graphical models that span the range of possible latent structures: from no to low-rank confounding, via intermediate regimes; and likewise for S_{ZX} .

The structure of the graph over X , S_X^* is identical for all simulations: it consists of a chain of length p in which one link out of five has been removed : $S_{X,ij}^* = S_{X,ji}^* = \rho$ if $j = i + 1$ and $i \not\equiv 0 \pmod{5}$. $S_{X,ij}^* = 0$, for all other pairs (i, j) . Also, $S_{X,ii}^* = 1$.

S_{ZX}^* is always constructed as follows. It is a matrix of size $2p \times p$ such that $S_{ZX}^* = [\beta \rho A_p; \zeta \rho B_p]$ (here “;” denotes vertical concatenation), where A_p, B_p are *random* matrices obtained by permuting the columns of the identity matrix I_p . The permutations are here to guarantee that the inputs and hidden variables are connected to a random subset of X . β and ζ control the strength of the effect of Z and H , respectively.

To account for the sparsity/low-rank structure of Z and H , we generate Y_Z and Y_H according to a simple tree model. We assume that p is an integer of the form $p = 2^k$; pick an integer $d_Z \in \{0, \dots, k\}$; and let $Y_Z^{(2^{d_Z})} \in \mathbb{R}^{2^{d_Z}}$ be a standard Gaussian random vector (*i.e.* the covariance matrix is the identity). Then, $Y_Z^{(2^{d_Z+1})}$ is defined inductively according to

$$Y_Z^{(2^{d_Z+1})} = \begin{pmatrix} Y_Z^{(2^{d_Z})} \\ Y_Z^{(2^{d_Z})} \end{pmatrix}.$$

We set $Y_Z = Y_Z^{(2^k)}$ and assume that only $Y_Z^{(2^{d_Z})}$ is observed. Y_H is defined according to the same process (replace d_Z with d_H).

If $d_Z = 0$ then a single input has an effect on all the variables in X . On the other hand, if $d_Z = k$, then there are k distinct inputs that are in a one-to-one correspondence with the variables in Z . If $d_Z = k - 1$, then each input acts on two

random nodes of X , thus the “effective” S_{ZX} is fairly sparse. By varying d_Z and d_H , we control the sparsity/low-dimensionality of the inputs and the latent variables. For example, whenever, $d_H < k$, hidden variables confound S_X .

From now on, we write $\mathcal{G}_{d_Z, d_H}$ for the graphical structure generated by choosing some pair (d_Z, d_H) . Figure 5.1 shows the structure of the graphs for different values of d_Z, d_H . As d_Z and d_H depart from k , the relationship between Z and X – along with the structure over X – become harder to estimate.

5.2 Results

In our simulations, we set $p = 2^5, \beta = \zeta = 5, \rho = 0.2$ and $n = 5 \times 10^5$. We then generate data according to $\mathcal{G}_{d_Z, d_H}$, letting d_Z and d_H range from 5 to 2. At one end of the spectrum ($d_Z = d_H = 5$) there is no confounding and the relationship between Z and X is easy to learn, while at the other end ($d_H = 2$) there are only a few confounders with an effect spread over many observed variables: this falls directly within the range of applicability of the low-rank plus sparse decomposition. The intermediate regime is where our method is more likely to be beneficial: the effect of Z is not so sparse, confounding is not so incoherent. By generating data from these 16 models we try to provide a minimal set of designs that span the range of situations one might face in real applications.

The performance of all four methods (graphical lasso, sparse conditional GGM, low-rank plus sparse, our method) across the 16 simulation designs is shown in Figure 5.2. Here, we are interested in recovering the structure of S_X and use precision / recall curves as a metric, thus ignoring the rank of the latent component.

LR+S and LSCGGM both have two tuning parameters (λ and γ). For each value of γ , one obtains a different precision/recall curve by varying λ . Figure 5.2 shows the curve obtained by selecting γ in order to maximise recall at a precision of 1. Only

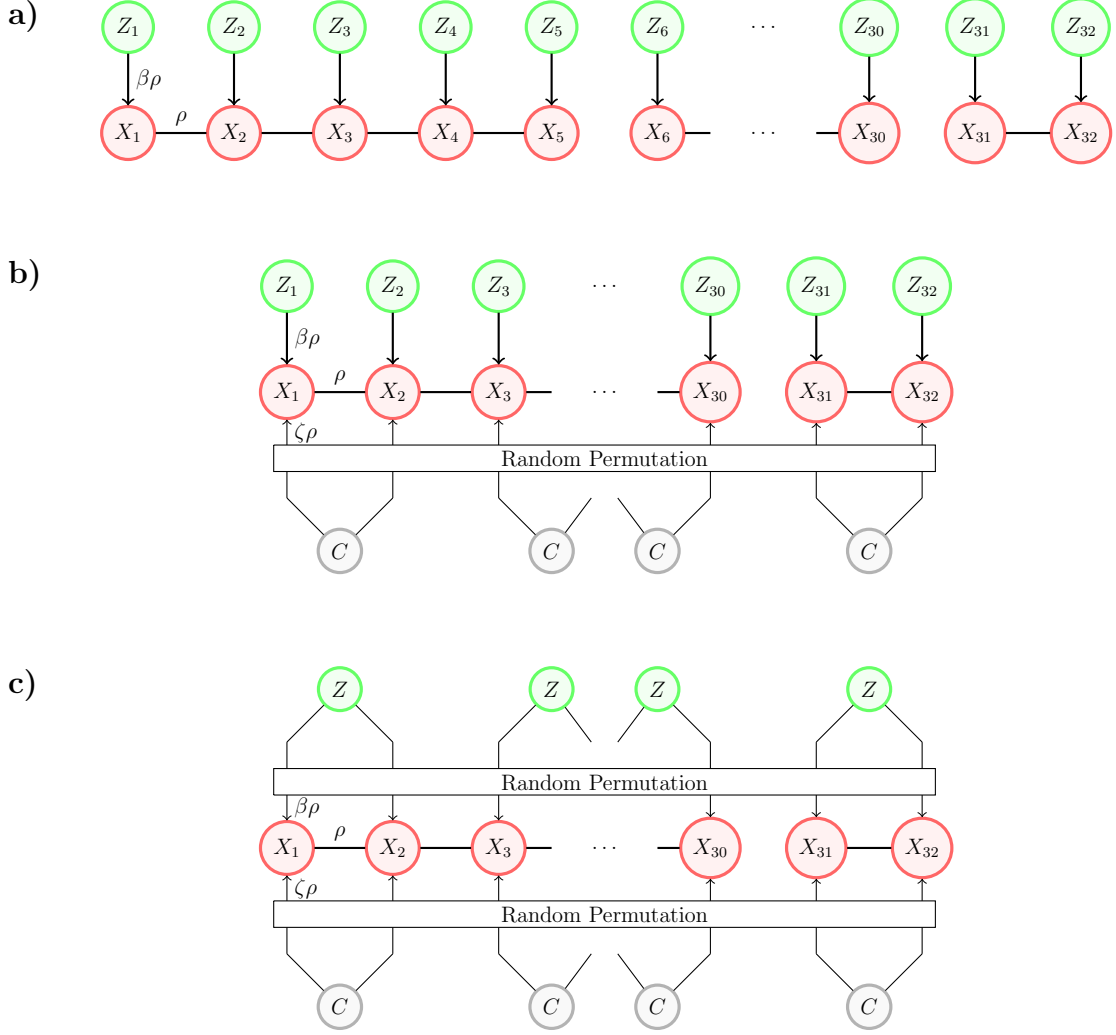


Figure 5.1: Graphical representation of the models described by $\mathcal{G}_{54}$ and $\mathcal{G}_{44}$, for $p = 32$. (a) $\mathcal{G}_{55}$. Here, $d_Z = 5, d_H = 5$. Each input acts on a single output. Hidden variables (not represented) act on a single output and, therefore, do not confound anything. (b) $\mathcal{G}_{54}$. Graph corresponding to $d_Z = 5, d_H = 4$. Each confounder (grey nodes) acts on two random observed variables (in red). With $p = 32$, there are 16 confounders. (c) $\mathcal{G}_{44}$. Graph corresponding to $d_Z = 4, d_H = 4$. Inputs (green nodes) act on two random observed variables.

results for $\rho = 0.2$ are shown here but similar experiments have been conducted for $\rho = 0.05, 0.1$ and gave similar results.

First, we see that known methods behave as expected. The graphical lasso achieves consistency when there is no confounding and Z acts in a sparse fashion. The sparse conditional GGM approach is more robust to changes in S_{ZX} but this is restricted to situations in which there is no confounding. The low-rank plus sparse (LR+S) method achieves perfect recovery when there are few latent variables (four, for a total of 32 observed variables); this corresponds to $d_H = 2$. Whenever any of these three methods recovers the graph perfectly, our method (LSCGGM) offers comparable performances. $d_H = 4$ corresponds to the extreme situation in which each latent variable confounds exactly two random variables.

More importantly, we identify a regime ($d_H = 3$) in which LSCGGM outperforms the other methods. In this regime, there are 8 latent variables (each latent variable affects 4 random observed nodes), so that the effect of the hidden variables is not so incoherent. Such a regime is far less restrictive than the one required for the consistency of the LRPS method, thus showing that we benefit from the additional assumptions made about the data.

As mentioned above, both LR+S and the suggested approach (LSCGGM) have two tuning parameters: λ and γ . While λ controls the overall shrinkage on the sparsity/rank of the estimates, γ accounts for the trade-off between sparse and low-rank components. In order to understand the effect of γ on the regularisation paths of these methods, we study the surface formed by the set of precision/recall curves obtained for various values of γ . Following the suggestion made in Chandrasekaran et al. (2009) we parametrise the penalty term as $\lambda(\gamma\|S\|_1 + (1 - \gamma)\|L\|_*)$, so that γ ranges from 0 to 1 instead of $(0, +\infty)$. Figure 5.3 illustrates how precision/recall curves evolve as one changes the relative weight given to the $\|\cdot\|_1 / \|\cdot\|_*$ norm. By analogy to the Area Under Curve (AUC) metric, we measure the “Volume Under

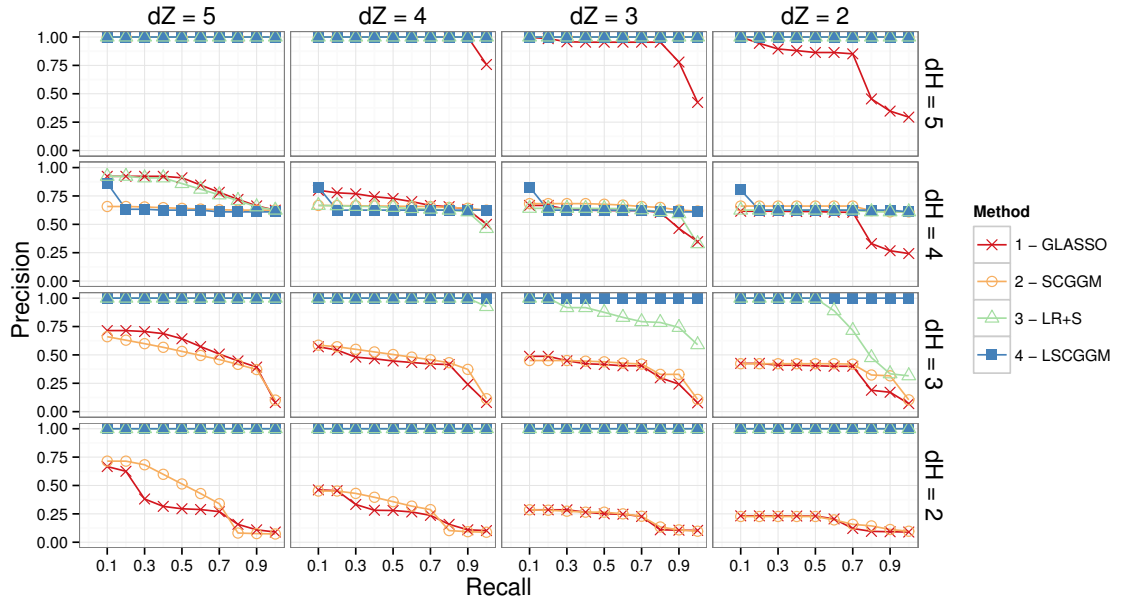


Figure 5.2: Comparison of the suggested estimator (LSCGGM) to other published methods. Along the x -axis (resp y -axis), d_Z (resp d_H) varies from 5 to 2. More precisely, in the top row, there is no confounding at all. In the second row from the top, hidden variables act in a very sparse fashion. In the bottom row, there are 4 hidden variables and we are in the range of applicability of the low-rank plus sparse method. The third row corresponds to an intermediate regime in which there are twice as many latent variables. Settings: $p = 2^5$, $n = 5 \times 10^5$, $\beta = \gamma = 5$, $\rho = 0.2$. Each design is repeated 25 times. The value of the tuning parameter γ was chosen to maximise recall at a precision of 1.

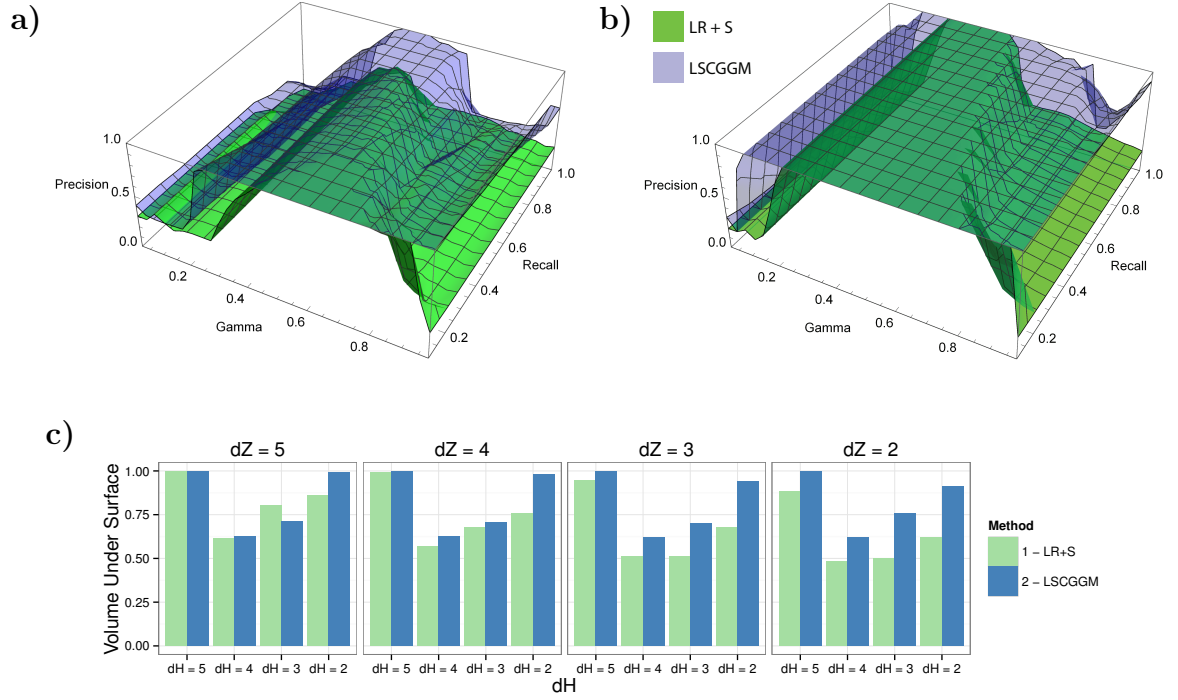


Figure 5.3: Sensitivity to the tuning parameter γ . Here, an alternative parametrisation of the regularisation term is used: $\lambda_n(\gamma\|S\|_1 + (1 - \gamma)\|L\|_*)$, so that $\gamma \in (0, 1)$ instead of $(0, +\infty)$. (a) Precision/recall surface for $\mathcal{G}_{23}$ (*i.e.* each input Z acts on 8 random outputs and there are 8 confounding variables) (b) Precision/recall surface for $\mathcal{G}_{33}$ (there are 4 confounding variables). (c) Volume under surface across all 16 simulation designs.

Surface” which accounts for the effect of both regularisation parameters.

Figure 5.3 provides an illustration of some of the theoretical results of Chapter 3: when achieved, consistency holds for a range of values of γ . This is also what was observed in the simulations of the previous chapter. Furthermore, for almost all designs, this range is greater for the LSCGGM method than for the LR+S approach, thus reducing the sensitivity to this tuning parameter.

Chapter 6

Application: Using Genetic Information to Detect Relationships between Human Metabolites

We now apply the suggested method to a dataset combining human metabolite levels and genetic markers. Here, metabolites play the role of the variables indexed by X while genetic variants are used as the inputs, Z . For comparison purposes, we also report the results obtained with the Low-Rank plus Sparse method (LR+S)¹.

6.1 The Avon Longitudinal Study of Parents and Children (ALSPAC)

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a cohort study of children born in the county of Avon during 1991 and 1992 (Pembrey 2004).

¹The other two methods (graphical lasso and sparse conditional graphical model) arise as special cases by setting γ close to 0.

For obvious privacy reasons this dataset is not publicly available but the data can be requested by submitting a research proposal at the following address: [`https://proposals.epi.bristol.ac.uk/`](https://proposals.epi.bristol.ac.uk/)².

The data at our disposal contained genetic and phenotypic measurements on approximately 8,000 children and their mothers. We identified around sixty metabolic traits of interest and kept the variables for which a sample size of 2,000 or more was available in both cohorts. 39 traits passed this filter and are listed in the table below (Table 6.1). Measurements for all 39 variables were available without missing data for 5,242 children and 2,770 mothers.

For every sample, hundreds of thousands of genetic markers were measured. We therefore screened genetic variants in order to include only those markers that had a good predictive power for at least one of the 39 traits.

This was achieved as follows. For every of the 39 metabolic traits, we performed independent univariate linear regressions of the trait on each of the genetic variants. Thus, for each trait, we obtained hundreds of thousands of p-values that we corrected for multiple testing using a Benjamini-Hochberg procedure. All variants that were significant at a False Discovery Rate of 10% were selected. In each cohort, a few hundred markers were significant at this FDR level but many of them were strongly correlated due to linkage disequilibrium (variants that are next to each other in the genome are typically very strongly correlated but often tag a single causal variant.) We selected variants so that their Pearson correlation be at most 0.2. 133 and 44 variants passed this filter in the child and mother cohorts respectively.

Metabolite levels being continuous variables, they were quantile normalised and standardised. Genotypes, on the other hand, were encoded as ternary variables (0/1/2).

²Please note that the study website contains details of all the data that is available through a fully searchable data dictionary (<http://www.bristol.ac.uk/alspac/researchers/access/>). Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

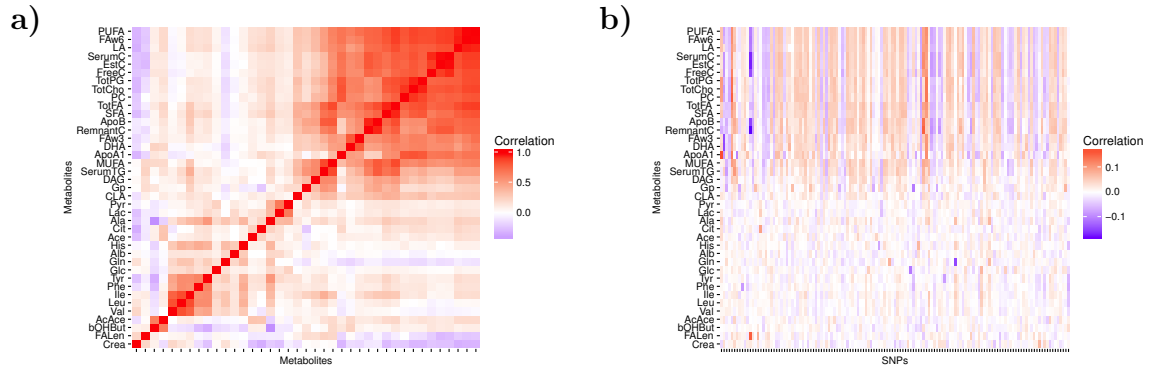


Figure 6.1: a) Correlation matrix between phenotypes (X) b) Correlation matrix between metabolites (X) and SNPs (Z). These are computed on the Child dataset.

In summary, for the Child cohort (resp. Mother cohort) we have: $n = 5242$, $p = |X| = 39$, $m = |Z| = 133$ (resp. $n = 2770$, $p = 39$, $m = 44$).

6.1.1 Relevance of the ALSPAC dataset

We believe that the ALSPAC dataset has features that make it particularly interesting as an application dataset: large sample size compared to the number of variables, independent cohorts, continuous traits, etc... Perhaps the most interesting property of this dataset is related to the mode of action of the genetics (variables in Z) on the observed phenotypes. As opposed to low-level phenotypes such as gene expression levels - most of whose variance can often be explained by a single variant - metabolite levels can be impacted by many different variants through a number of complex pathways. Furthermore, as illustrated by Figure 6.1, most metabolites have a common genetic basis: a single genetic variant can have an impact on multiple traits. It is thus clear that modelling the relationships between Z and X using a sparse matrix is not well-suited to this kind of dataset. On the other hand, assuming that the dense correlation matrix depicted in Figure 6.1b can be explained by a handful of latent variables and a few sparse connections seems more appropriate.

6.2 Methods

6.2.1 Selection of the tuning parameters

In order to select the tuning parameters (λ_n, γ) , we follow the procedure described in Chapter 4. We consider 30 values of γ within the range $[0.02, 0.98]^3$. To each γ corresponds a regularisation path: a graph along each path is selected using “point-wise” complementary pairs stability selection (Meinshausen & Bühlmann 2010, Shah & Samworth 2013) (here $B = 100$, so that inclusion probabilities are computed over 100 folds, see Section 4.2.1 in Chapter 4 for more details.). Following the approach used in Meinshausen & Bühlmann (2010), the threshold on the inclusion probabilities is chosen by requiring that the expected number of falsely discovered edges be at most one: $E(V) \leq 1$ (using the notation of Chapter 4). Thus, for each method and each cohort we obtain a collection of thirty graphical structures.

As before, we measure how similar two graphical structures are, by considering their edge sets. For any pair of undirected graphs $\mathcal{G}_1 = (V_1, E_1)$, $\mathcal{G}_2 = (V_2, E_2)$ we define their similarity by their Jaccard Index

$$J(\mathcal{G}_1, \mathcal{G}_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}.$$

This measure has two uses: 1) it makes it possible to select γ by measuring how the estimates relate to each other as γ varies from 0 to 1; 2) it allows us to measure how well the findings are replicated across cohorts.

6.2.2 Assessing the biological relevance of the estimates

In order to evaluate the performance of the different methods, we computed the enrichment statistic described in Chapter 4, section 4.3. Here, we used *ChEBI: an*

³Recall that the penalty is parametrised as $\lambda(\gamma\|S\|_1 + (1 - \gamma)\|L\|_*)$, so that $\gamma \in (0, 1)$.

ontology of small chemical entities of biological interest (Hastings et al. 2012). All 39 nodes were manually matched to their ChEBI IDs. To each node, we associated a set of ChEBI IDs consisting of its own ID, “related” terms’ IDs and it’s parents IDs (related terms and parents are defined by the ChEBI ontology). Once the graphs were annotated, computing their scores was a straightforward application of (4.8).

6.3 Results

First, we can ask how sensitive the estimates are to the tuning parameter γ . Indeed, as pointed out earlier, one would expect to see a “stable region”: a range of values of γ for which there is little variation. One would typically select a graph estimated with a γ within this region. Let $\hat{\mathcal{G}}_{LSCGGM,Ch}^{(\gamma)}$ (resp. $\hat{\mathcal{G}}_{LR+S,Ch}^{(\gamma)}$) denote the graph returned by LSCGGM (resp. LR+S) for a given value of γ in the Child cohort. For every pair (γ_1, γ_2) , Figure 6.2 **a**) shows how similar the estimates are to each other (as computed by $J(\hat{\mathcal{G}}_{LSCGGM,Ch}^{(\gamma_1)}, \hat{\mathcal{G}}_{LSCGGM,Ch}^{(\gamma_2)})$). In the range $0.6 \leq \gamma_1, \gamma_2 \leq 0.9$, they are very close to each other. For small values of γ , the graphical structures returned by LSCGGM vary smoothly with γ . The regime $\gamma \leq 0.05$ corresponds to the case in which the rank of the latent component is 0: LSCGGM behaves like a sparse conditional graphical model. Figures 6.2 **b**), **d**), **e**) show that LR+S behaves similarly, whether it is in the Child or the Mother cohorts.

Having established that both methods exhibit a stable region, we look at how close the estimates found in these regions are. To that end, we plot $J(\hat{\mathcal{G}}_{LSCGGM,Ch}^{(\gamma_1)}, \hat{\mathcal{G}}_{LR+S,Ch}^{(\gamma_2)})$ for all pairs (γ_1, γ_2) (see Figure 6.2 **c**)). For small values of γ , LR+S and LSCGGM seem indistinguishable. However, for $\gamma_1, \gamma_2 > 0.5$ their Jaccard Index drops to reach values around 0.3 - 0.4. But the range $\gamma > 0.5$ covers precisely the stable regions of both LSCGGM and LR+S, thus indicating that the methods’ “best guesses” are different. We observe the same behaviour in the Mother cohort (Figure 6.2 **f**)). Figure

6.3 shows in what way the graphs found in those stable regions differ, with LR+S inferring more connections between amino-acids (*e.g.* alanine, valine, leucine, etc... as opposed to lipids: cholines, cholesterol, etc...).

Given that two cohorts are at our disposal, one way of assessing the quality of our results is to look at how well they replicate across datasets. In Figure 6.4 **a)** we plot the similarity between graphs estimated at the same value of γ (Figure 6.4 **b)**, **c)** plots this similarity for all possible pairs γ_1, γ_2). First, it can be seen that higher replication values are achieved in the stable regions of their respective methods, with Jaccard Indexes at 0.6 or above. We also see that LSCGGM’s edge set replicates better than LR+P’s. Moreover, the suggested estimator retrieves more edges under the condition $\mathbb{E}(V) \leq 1$ (see Figure 6.4 **d)**).

Finally, we use the “enrichment statistic” defined earlier (in section 4.3 of Chapter 4). In our attempt to assess the quality of our estimates and their biological relevance, this metric is useful as it makes it possible to score graphs using an external source of information. Figure 6.5 **a)** shows the value taken by this statistic across cohorts and methods. Associated p-values can be found in Figure 6.5 **b)**. Here again it is clear that, irrespective of the dataset, higher values are achieved within the stable regions of their respective methods. Just like in the case of the replication measure, LSCGGM achieves the highest values. Given that the Child cohort contains twice as many samples as the Mother cohort, it is surprising to observe better performances in the Mother dataset. This might be due to the fact that this cohort is more homogeneous: there are women only, measurements were taken the same number of months after pregnancy, etc...

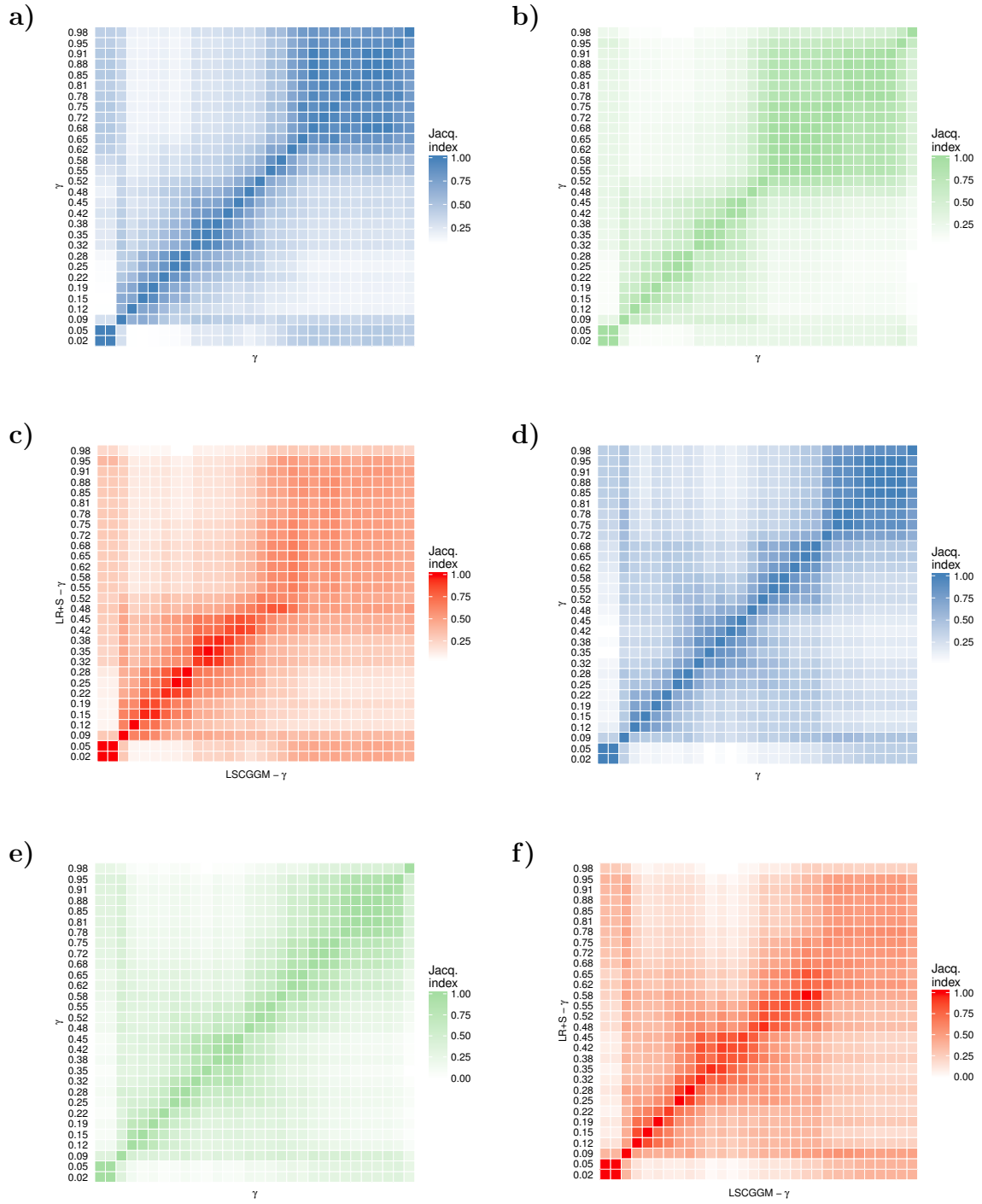


Figure 6.2: Sensitivity of LSCGGM and LR+S to the tuning parameter γ . For any two graphs, their similarity is computed using the Jaccard Index of their edge sets. (a) Similarities between the edges sets of the graphs returned by LSCGGM in the *Child* cohort, as a function of γ (for 30 values of $\gamma \in (0.02, 0.98)$). (b) Same as (a), but with L+PS. (c) Similarities between the graphs returned by LSCGGM and LR+S in the *Child* cohort. (d) Same as (a), in the *Mother* cohort. (e) Same as (b), in the *Mother* cohort. (f) Same as (c), in the *Mother* cohort.

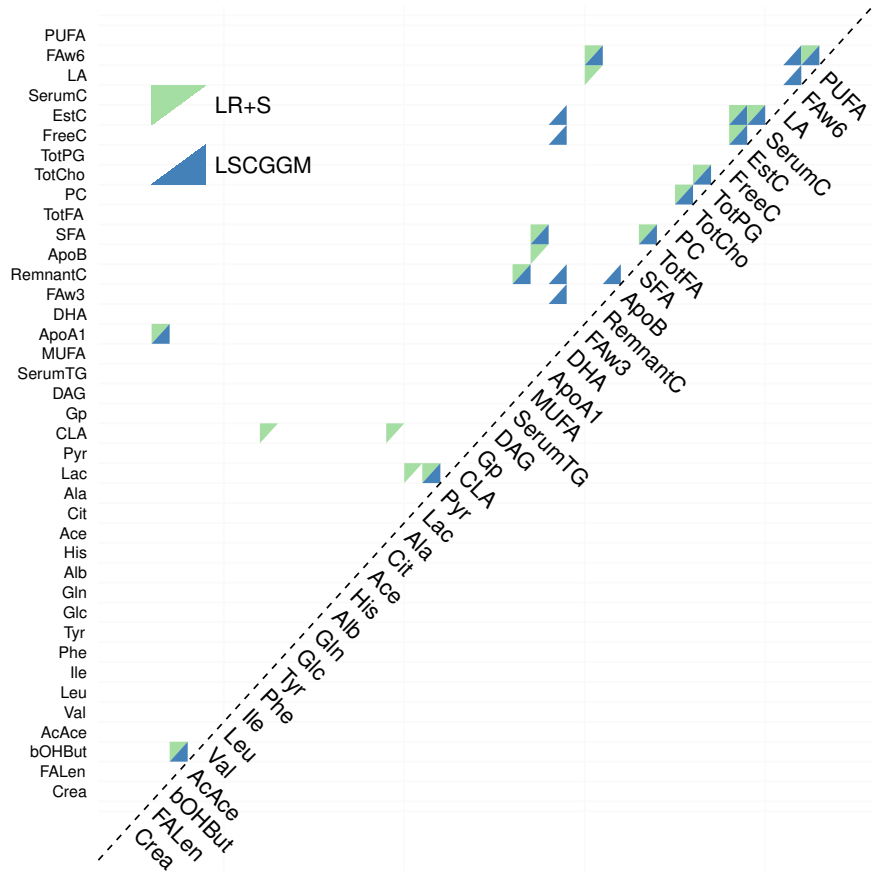


Figure 6.3: Adjacency matrices of the graphs returned by the LSCGGM and LR+S methods for $\gamma = 0.81$ and $\gamma = 0.68$ respectively.

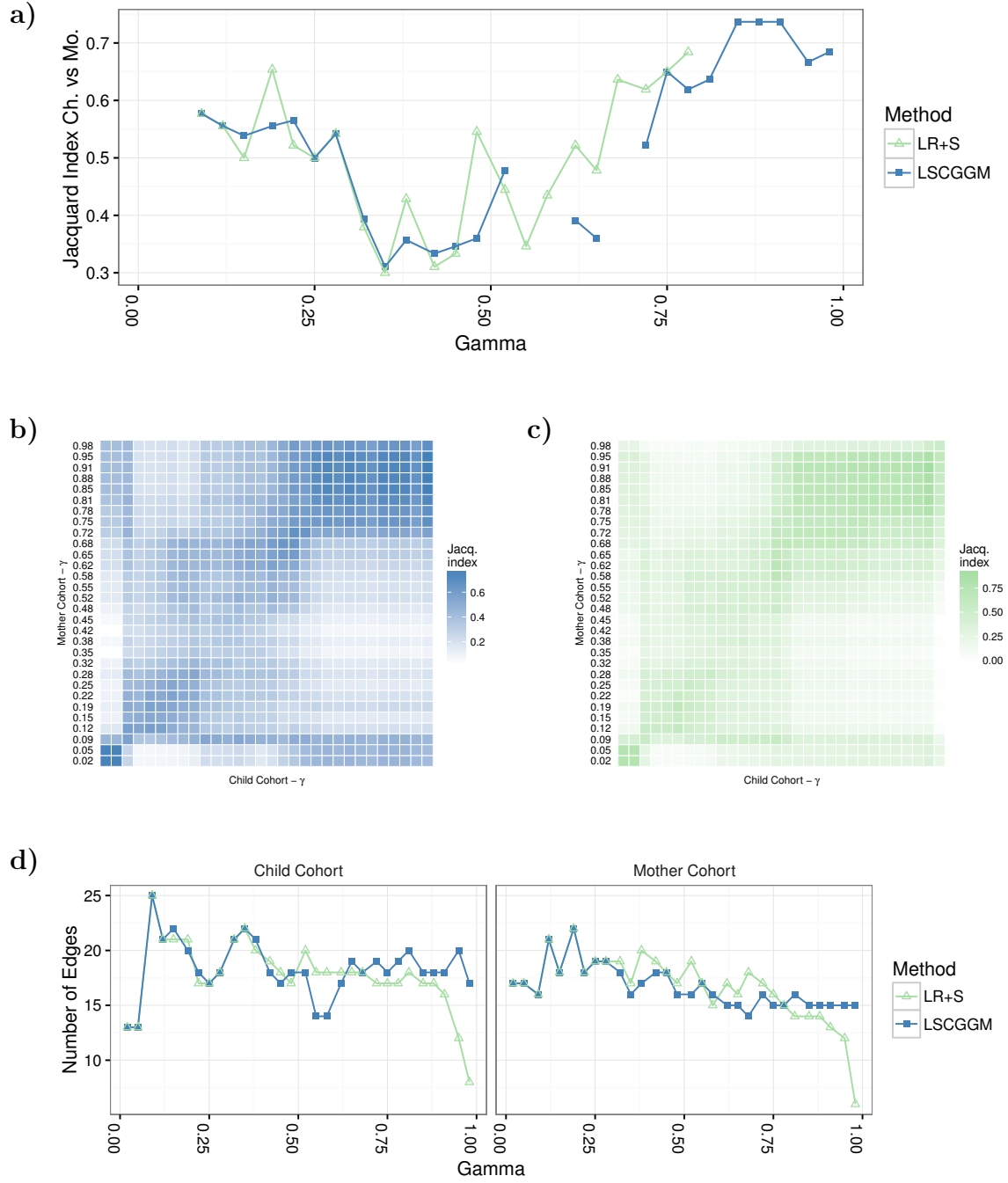


Figure 6.4: a) Comparing estimates across cohorts. For each value of γ and each method, we plot the similarity between the estimate obtained in one cohort against the one obtained in the other. We limit ourselves to values of γ for which the estimates in *both* cohorts comport 15 edges or more. b) Similarities across cohorts for all possible pairs (γ_1, γ_2) , for LSCGGM. c) Same as b), with LR+S. d) Number of edges retrieved by LR+S and LSCGGM as a function of γ in both cohorts under the condition $\mathbb{E}(V) \leq 1$.

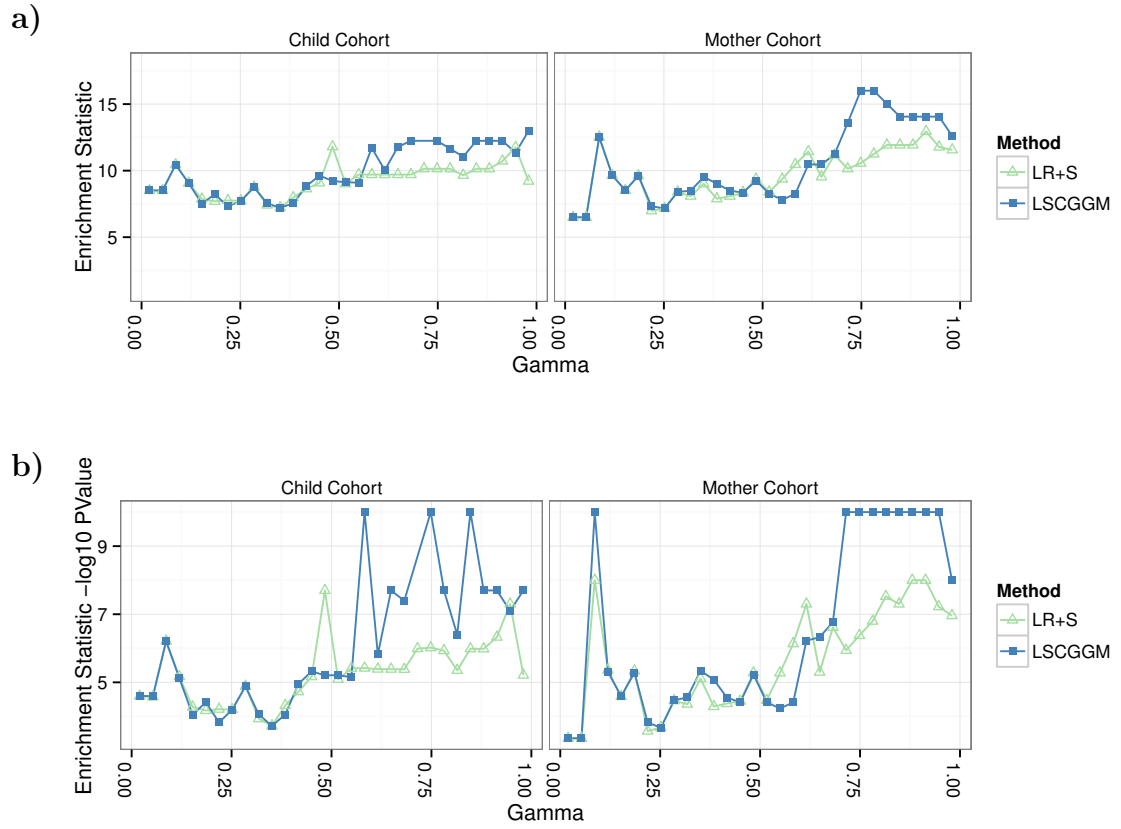


Figure 6.5: a) Enrichment statistic, as a function of the tuning parameter γ . b) p-value of the enrichment statistic, as a function of γ . The p-values were computed from 1 billion permutations. In some cases, none of the permuted statistics were greater or equal than the statistic of the actual estimates, hence the saturation.

6.4 Discussion

In this chapter, we analysed a well-powered dataset comprising genetic and metabolite measurements. A lot of these measurements were conceptually very close to each other. For example, we modelled the levels of polyunsaturated fatty acids (PUFA) and $\omega - 3, 6$ fatty acids (FAw6, FAw3). But $\omega - 3, 6$ fatty acids *are* polyunsaturated fatty acids and, in turn, linoleic acids (LA) *are* FAw6 (all these relationships were correctly identified by LSCGGM). Moreover, many loci impact the levels of many different traits, thus making it probable that these correlations can be further explained by latent pathways. These features, combined with the large sample size made this dataset an interesting application: 1) we knew that some of the relationships were causal and; 2) we knew that it was challenging to uncover them due to confounding; 3) we had enough samples to estimate a complex latent structure (high-value of γ).

Table 6.1: Descriptions and symbols of the 39 metabolites modelled in this chapter.

Variable	Description
serumc	serum total cholesterol (mmol/l)
remnantc	remnant cholesterol (non-hdl non-ldl -cholesterol) (mmol/l)
estc	esterified cholesterol (mmol/l)
freec	free cholesterol (mmol/l)
serumtg	serum total triglycerides (mmol/l)
dag	diacylglycerol (mmol/l)
totpg	total phosphoglycerides (mmol/l)
pc	phosphatidylcholine and other cholines (mmol/l)
totcho	total cholines (mmol/l)
apoal	apolipoprotein a-i (g/l)
apob	apolipoprotein b (g/l)
totfa	total fatty acids (mmol/l)
falen	estimated description of fatty acid chain length not actual carbon number
dha	22:6-docosahexaenoic acid (mmol/l)
la	18:2-linoleic acid (mmol/l)
cla	conjugated linoleic acid (mmol/l)
faw3	omega-3 fatty acids (mmol/l)
faw6	omega-6 fatty acids (mmol/l)
pufa	polyunsaturated fatty acids (mmol/l)
mufa	monounsaturated fatty acids; 16:1 18:1 (mmol/l)
sfa	saturated fatty acids (mmol/l)
glc	glucose (mmol/l)
lac	lactate (mmol/l)
pyr	pyruvate (mmol/l)
cit	citrate (mmol/l)
ala	alanine (mmol/l)
gln	glutamine (mmol/l)
his	histidine (mmol/l)
ile	isoleucine (mmol/l)
leu	leucine (mmol/l)
val	valine (mmol/l)
phe	phenylalanine (mmol/l)
tyr	tyrosine (mmol/l)
ace	acetate (mmol/l)
acace	acetoacetate (mmol/l)
bohbut	3-hydroxybutyrate (mmol/l)
crea	creatinine (mmol/l)
alb	albumin (signal area)
gp	glycoprotein acetyls mainly a1-acid glycoprotein (mmol/l)

Chapter 7

Application: Using Expression Quantitative Trait Loci to Model the Expression Levels of Human Genes

In Chapter 6, we applied our method to a dataset containing dozens of variables and thousands of samples. This application was interesting because, due to the nature of the variables being modelled, we knew that latent variables would be mediating the effect of the genetic variants. Moreover, thanks to the large number of samples available, we could afford to learn a relatively complex unobserved structure: both our method and the low-rank plus sparse method were well-behaved for high values of γ ($\gamma > 0.8$) – *i.e.* when most of the penalty is on the sparse component.

In the present chapter, we consider a dataset which is, in many respects, very different from the one studied previously: most inputs are in a one-to-one mapping with the outputs, we expect few latent variables and the number of variables is greater than the number of samples. The application described here is therefore complementary

to the one given earlier.

Here, we seek to model the expression levels of hundreds of human genes (X) conditional on a number of genetic variants (Z) referred to as *expression Quantitative Trait Loci* (eQTLs). This dataset was first analysed in Battle et al. (2013). It was generated by measuring the expression of the genes of 922 individuals, in whole blood.

To better understand the problem at hand, we start by recalling some relevant elements of biology. We then describe the dataset under study and briefly go through the data preparation process and methods. We conclude with our results.

7.1 Expression Quantitative Trait Loci (eQTLs)

The data we study here is an *RNA-seq gene expression dataset*. To understand how such data is generated, recall that the genetic code contained in the DNA encodes instructions on how to make proteins. According to the central dogma of biology, two main steps are required in order to convert the information contained in the DNA into proteins:

1. *transcription* (from DNA to mRNA): this is the process by which the information contained in a piece of DNA is converted into mRNA (messenger RNA).
2. *translation* (from mRNA to protein): this is the process by which an mRNA is converted into a protein by the ribosome.

The dataset at our disposal was generated by sequencing the mRNA contained in cells extracted from whole human blood. By doing so, it is possible to quantify the amount of mRNA produced by each of the genes of a given subject, at a given moment in time. Looking at the expression levels of hundreds of people, one can hope to uncover co-expression patterns, interactions between genes and even get some understanding of the causal mechanisms underpinning higher-level traits and diseases.

As useful as RNA-seq is, it is important to keep in mind that it only provides a proxy for protein levels. Indeed, in both steps (transcription and translation) some information is lost: some mRNAs are never translated into proteins, some mRNAs are reused, etc.... Thus, changes in the levels of some mRNA might never impact the levels of the corresponding protein. When genes interact with each other (*i.e.* up-regulate or down-regulate each other) this usually happens because a protein – not an mRNA – increases the amount of mRNA made for another protein. That being said, we can still expect some of the signal found in an RNA-seq dataset to be propagated and to reflect actual gene-gene interactions. For example, a number of loci act on both their nearby gene (cis-gene) and distant genes, making it plausible that such loci affect their cis-gene which, in turn, affects the levels of the remote gene (the precise definition of cis-gene is given hereafter) (Westra et al. 2013).

In some RNA-seq datasets, genetic information is also available. eQTLs (expression Quantitative Trait Loci) are genomic loci that regulate the expression of mRNAs. Depending on whether they act on the expression of a nearby gene or remotely, eQTLs are divided into *cis-eQTLs* and *trans-eQTLs* respectively. In order to identify cis-eQTLs, one usually tests whether any of the loci situated within 1Mbp of a given gene are significantly correlated with the expression levels of that gene ¹. It has been observed that most genes have at least one cis-eQTL (Aguet et al. 2016). Furthermore, many cis-eQTLs are very good predictors of the levels of their cis-gene (as will be shown in the next section).

Due to a phenomenon called *linkage disequilibrium*, genetic variants that are close to each other in the genome tend to be strongly correlated (and sometimes perfectly correlated). This might create some issues when computing the co-expression of two genes located in the same region of the same chromosome. The expression of two such genes might be confounded by some unobserved, or unaccounted for, genetic

¹Note that this 1 million base pairs cut-off is arbitrary and some will prefer using a 500Kbp window.

variant. When applying our method, we typically do not include all of the variants at our disposal but rather a few highly predictive loci. By doing so, we allow linkage disequilibrium to confound our analysis. Consequently, all intra-chromosomal edges will be discarded (edges of the final estimate which connect two genes that are on the same chromosome).

7.2 Data Preparation

RNA-seq data is prone to confounding for a number of reasons. For example, if the samples have been sequenced in different facilities (or at different times of the year, or at different temperatures, etc...) there can be “batch effects” that will eventually lead to false discoveries. When trying to identify eQTLs, such confounders must therefore be accounted for (some of these are observed and can be included as covariates). Processing raw RNA-seq data is challenging and requires a good understanding of the data generation process. Thankfully, the dataset we acquired had already been analysed in Battle et al. (2013) and the data obtained from the NIH could be readily used to replicate their results.

After a first pass of processing of the raw data, the authors of Battle et al. (2013) regressed the expression of all the genes on so-called *technical factors* (*e.g.* time of the day) and kept the residuals of that regression. Then, for any given gene and variant, the residuals were used to test whether that variant was an eQTL for that gene. In this second step, additional *covariates* were also added to the model in an attempt to correct for some known and unknown confounders (the unknown confounders were estimated using a method called *Hidden Confounders with Prior* and designed specifically for RNA-Seq data (Mostafavi et al. 2013)).

The analysis presented in this chapter differs slightly from the one of Battle et al. (2013) because we regressed out both technical factors and covariates in the same

step (using ordinary least squares). We then quantile normalised the residuals and standardised them. As usual, genetic variants were initially encoded as vectors taking values in $\{0, 1, 2\}$ and then standardised.

At this stage, our dataset contained the expression of 15,230 genes along with the genotypes of approximately 1 million genetic variants, for a sample size n of 922. Both our current solver and the published solver for the low-rank plus sparse method are incapable of dealing with such a large number of variables. We therefore decided to keep only those genes and variants that were involved in the strongest cis-variant/cis-gene associations. In their supplementary materials, the authors of Battle et al. (2013) reported the summary statistics of their cis-eQTL findings. We used this information to rank variant/genes pairs according to the strength of their association (as measured by the p-value of the test performed in Battle et al. (2013))². Of all 10,914 independent cis-eQTLs identified in Battle et al. (2013), the 800 strongest variant/gene pairs were selected. Because some genes have multiple cis-eQTLs and because some variants are eQTLs for more than one gene, keeping the 800 most significant associations led us to select 789 genes and 774 variants. As mentioned earlier, some variants can be strongly correlated with each other. In order to avoid singularities, we removed variants that had a Pearson correlation with another variant greater than 0.2: 622 loci passed that filter. In summary, the data studied here has the following properties: $|X| = 789$, $|Z| = 622$ and $n = 922$.

To get a sense of how predictive the inputs Z are for the genes X , we performed a univariate linear regression of each gene on each variant. For every locus, we recorded the F-Statistic of the gene that was best predicted by that locus. Figure 7.1 **a)** plots the resulting distribution of F-Statistics. As can be seen from this Figure, most eQTLs are very predictive of at least one gene in X (namely, their cis-gene).

²Note that Battle et al. (2013) report only one eQTL per linkage disequilibrium (LD) block: they analysed the data further in order to identify which of the variants in a given LD block was most likely to be the causal one.

For comparison, recall that it is common in the instrumental variable literature to call an instrument *strong* if its F-Statistic is greater than 10 (Stock & Yogo 2002). Thus, an overwhelming proportion of the variants has an F-Statistic which is orders of magnitude greater than what is usually required.

Figure 7.1 **b)** shows that most of the variants act on only a few genes. Note that the correlation matrix plotted in Figure 7.1 **b)** is in sharp contrast with the one shown in Figure 6.1 in Chapter 6. trans-eQTLs being very rare, the diagonal pattern of the correlation matrix is not surprising: most eQTLs act only on their cis-genes while few of them act on an additional distant gene.

7.3 Methods and choice of the tuning parameters

As stated above, we are modelling the levels of hundreds of genes – 789 – so that the set of edges over which selection is performed has a cardinality in the hundreds of thousands (310,866). Moreover, since the sample size is 922, it is reasonable to expect the final estimates to contain hundreds of edges. With that many edges, we can afford to compare methods by comparing their regularisation paths – and not limit ourselves to a single estimate. Here, we vary λ within a reasonable range and record the resulting graphical structures as edges enter the model. To get a sense of what constitutes a “reasonable” range for λ , we used complementary pairs stability selection with the graphical lasso, under the condition $\mathbb{E}(V) \leq 10$ (according to the notations of Chapter 4). Here $\mathbb{E}(V) \leq 10$ was chosen instead of $\mathbb{E}(V) \leq 1$ because hundreds of stable edges are expected. The expected number of falsely discovered edges was therefore increased accordingly. This resulted in a graph with approximately 250 edges (after removal of approximately 60 intra-chromosomal edges). Consequently, we selected a range of values of λ that yielded between 10 and 600 inter-chromosomal edges.

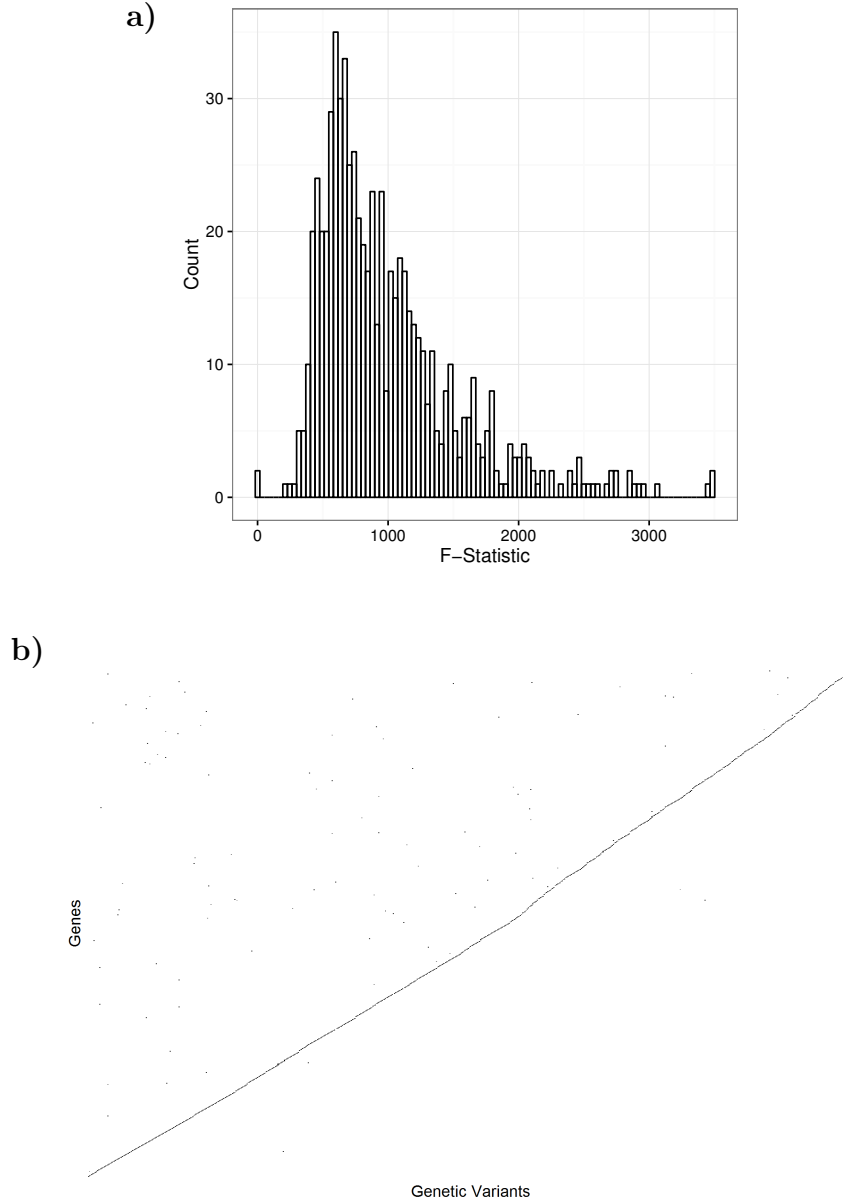


Figure 7.1: **a)** Distribution of F-Statistics across all 622 loci ($n = 922$, degrees of freedom of the F-Statistic: 920). For each genetic variant, we kept the F-Statistic for the gene that yielded the strongest association with that variant. Min F-Statistic = 13.4. **b)** Entries of the correlation matrix between genes and genetic variants with an absolute value greater than 0.2. For clarity, all the entries with an absolute correlation less than 0.2 were set to 0, while all the entries with an absolute correlation greater than 0.2 were set to 1.

The methods compared here are the graphical lasso (GLASSO), the low-rank plus sparse method (LRPS) and the approach suggested in this thesis (LSCGGM). GLASSO is obtained as a special case of LRPS by setting γ to a small value (0.01, in this instance). Likewise, the sparse conditional graphical model (SCGGM) of Wytock & Kolter (2013) is obtained from LSCGGM by setting γ close to 0, which explains why we do not mention SCGGM in our results.

Choice of the tuning parameter γ

The dataset analysed here is high-dimensional, with the total number of variables $p+m$ being greater than the sample size n . Although $p+m$ is not considerably greater than n – as is typically the case in the high-dimensional literature – this feature still has an impact and makes this application different from the ones encountered in our simulations and in the previous chapter. The main challenge stems from the fact that, as λ gets smaller, more and more latent variables enter the model. But for every new latent variable, a whole new eigenvector must be estimated, which increases the number of free parameters by a few hundred. When performing model selection in the high-dimensional regime, the rank of the latent piece must therefore increase relatively slowly with the number of edges. If this is not the case (γ is too large), only a handful of edges can be recovered. In addition, the problem quickly becomes numerically non-identified and solvers do not converge. In the next chapter, we will discuss this issue further in light of the results of this analysis. In the present setting, we found that $\gamma \simeq 0.12$ was already too large and returned estimates with only a dozen of edges or so. For both methods, the range $\gamma \in (0, 0.08)$ seemed acceptable and we used 5 evenly spaced values within that interval.

Choice of the tuning parameter λ

As stated above, we compare methods by looking at their regularisation paths computed from a range of values of λ chosen so that there are between 10 and 600 extra-chromosomal edges. We repeat this operation for each of the five values of $\gamma \in \{0.01, 0.26, 0.041, 0.057, 0.072\}$ and for each of the two methods $M \in \{LRPS, LSCGGM\}$.

For a given method M and a given value of γ , we write $n_e(\lambda)_M^{(\gamma)}$ for the number of extra-chromosomal edges retrieved by method M with tuning parameters set to the values λ and γ . For each M and γ , we selected 100 values of λ evenly spread within an interval satisfying (approximately) $10 \leq n_e(\lambda)_M^{(\gamma)} \leq 600$. The following table summarises the ranges of values of λ that were used. Taking the intersection of all the intervals $[n_e(\lambda_{min})_M^{(\gamma)}, n_e(\lambda_{max})_M^{(\gamma)}]$ yields $[22, 598]$.

Method	γ	λ_{min}	λ_{max}	$n_e(\lambda_{min})_M^{(\gamma)}$	$n_e(\lambda_{max})_M^{(\gamma)}$
LRPS	0.01	15.5	30	654	14
LRPS	0.026	5	12	669	13
LRPS	0.041	2.1	7	598	14
LRPS	0.057	0.8	4.5	632	12
LRPS	0.072	0.31	2.5	623	14
LSCGGM	0.01	15	28	689	19
LSCGGM	0.026	4.5	11	705	19
LSCGGM	0.041	1.68	6	632	22
LSCGGM	0.057	0.5	3.6	678	19
LSCGGM	0.072	0.15	1.8	616	15

7.4 Assessing the quality of the estimates

Just like in the previous application chapter (Chapter 6), we will use the enrichment statistic introduced in Chapter 4 in order to assess the biological relevance of the estimates returned by LRPS and LSCGGM. We first describe how we would proceed

for a generic statistic whose value would be a monotonic function of biological relevance. We then give more details about how we use the enrichment statistic defined in Chapter 4.

7.4.1 General procedure

Let S be some statistic which can be computed for any unweighted, undirected graph over the 789 genes being modelled in this chapter. We write $S(n_e)_M^{(\gamma)}$ for the value taken by this statistic when applied to the estimate with n_e edges returned by method M . Given the choices made for λ , we obtain, for each method M and γ , 100 values of $S(n_e)_M^{(\gamma)}$ (with n_e in the range $10 \lesssim n_e \lesssim 600$). Finally, we get an estimate of the function $\mathbb{N} \rightarrow \mathbb{R}^+, n_e \mapsto S(n_e)_M^{(\gamma)}$ with a linear interpolation.

Thus, for every statistic S , method M and value of γ , we can plot S as a function of the number of edges and see how it changes along the regularisation path. It is also possible to compute the area under that curve (AUC), a quantity which is proportional to the average value of S over a given range. Depending on the statistic being used, the value of the AUC might be hard to interpret. Here, the graphical lasso (GLASSO) is our baseline method and all AUCs are rescaled by the AUC of the graphical lasso.

7.4.2 Main measure: the enrichment statistic of Chapter 4

We use the enrichment statistic $\mathcal{S}$ described in Chapter 4. The variables being genes, the relevant ontology is the Gene Ontology (GO). To be more precise, what is commonly referred to as the gene ontology is actually a collection of three ontologies named *biological process* (BP), *molecular function* (MF) and *cellular component* (CC) (Ashburner et al. 2000). We annotated the nodes of our estimates with the terms of all three ontologies and calculated the value (and associated p-value) of our enrichment statistic for each of the estimates obtained along the different paths (there is

path per value of γ and per method).

7.5 Results

First, Figure 7.2 **a)** and **b)** show how the AUC of our enrichment statistic (and its p-value) varies as a function of γ . The AUC was computed by integrating the curves over the interval $n_e \in [50, 500]$. As long as γ is not too large, we can see that the AUC for both methods is greater than 1, indicating that they outperform the graphical lasso. LSCGGM also offers a better average performance than LRPS. For both methods, the optimal γ appears to be 0.041.

In Figure 7.2 **c)** and **d)** we plot the curves obtained for $\gamma = 0.041$, with $n_e \in [50, 500]$. For n_e large enough, we see that LSCGGM outperforms LRPS, although the two curves are sometimes identical. The dip observed for LSCGGM at $n_e = 50$, can be explained by the fact that the curves are very unstable when there are few edges: they quickly jump from one extreme value to another. These plots indicate that LSCGGM tends to retrieve relevant edges earlier in the path compared to LRPS. It then starts including noise while LRPS slowly includes the same relevant edges. Thus, LSCGGM is capable of achieving a higher signal to noise ratio or, in other words, a better precision at equal recall.

Based on these curves, it is not possible to tell whether LRPS and LSCGGM tend to include the same relevant edges. Indeed, LRPS and LSCGGM could very well be sensitive to different kinds of signal. In order to rule that possibility out, we looked at the set differences and intersection of the edge sets returned by LRPS and LSCGGM for $n_e \simeq 200$ (and $\gamma = 0.041$). By doing so, it is possible to tell whether there is signal in the edges that are found in one method and not in the other. As can be concluded from Figure 7.3, the edges retrieved by LRPS but not LSCGGM appear to be noise. On the other hand, the edges that are specific to LSCGGM have a high

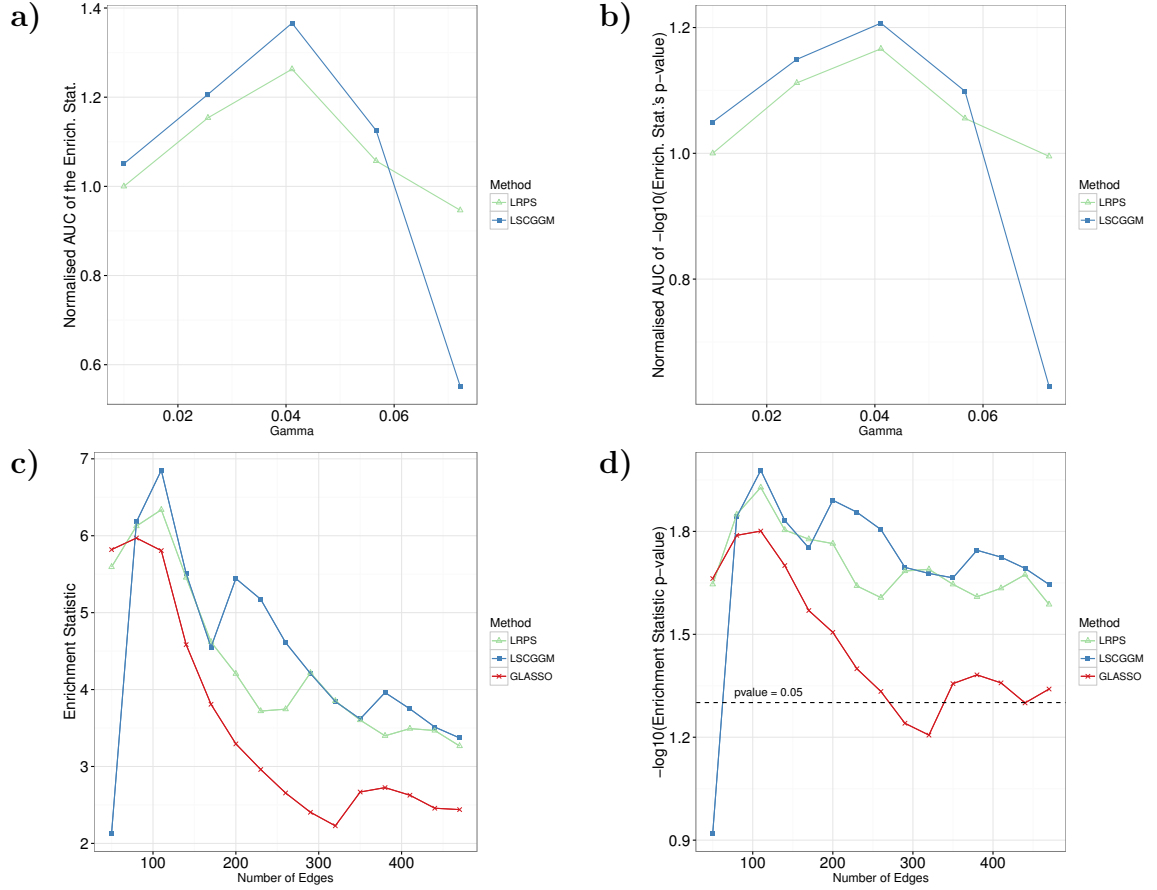


Figure 7.2: **a)** Area under the curve “Enrichment Statistic = $f(\text{Number of Edges})$ ” as a function of γ . The AUC is proportional to the average value of the function over this range. It is further normalised by the area under the curve of the graphical lasso, *i.e.* the area under the red curve in subplot c). **b)** Similar to a), but this time with the curve “ $-\log_{10}(\text{p-value of the enrichment statistic}) = f(\text{Number of Edges})$ ”. The AUCs are normalised by the AUC of the red curve in subplot d). **c)** Value of the enrichment statistic as a function of the number of edges, for $\gamma = 0.041$. **d)** Value of the $-\log_{10}(\text{p-value})$ associated with the enrichment statistic as a function of the number of edges, for $\gamma = 0.041$. All p-values are calculated from 10,000 permutations.

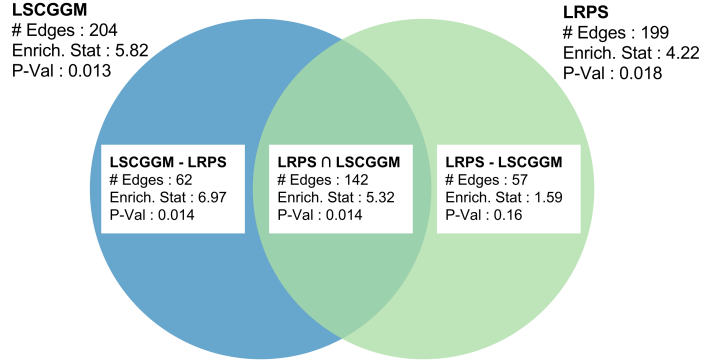


Figure 7.3: Venn diagram of the edge sets returned by LRPS and LSCGGM, for values of λ returning approximately 200 edges. At the top left corner (resp. right corner), the number of edges, score and p-value of the LSCGGM estimate (resp. LRPS). The left hand side of the diagram corresponds to the edges that are retrieved by LSCGGM but not LRPS. The middle portion corresponds to those edges that are retrieved by both methods.

score. Performing the same analysis for $n_e = 500$, a number of edges for which the graphs have similar scores, revealed that all of the signal contained in either of the estimates was concentrated in their intersection.

We further investigated the properties of the estimates by looking at which Gene Ontology (GO) terms were significantly enriched. Just like in the previous paragraph, we first selected the graphs so that $n_e \simeq 200$ (and $\gamma = 0.041$, for LRPS and LSCGGM). We then looked at the GO terms that achieved an enrichment of 20 or greater:

- GLASSO (204 edges): 21 terms had an enrichment of 20 or greater;
- LRPS (199 edges): 35 terms had an enrichment of 20 or greater;
- LSCGGM (204 edges): 37 terms had an enrichment of 20 or greater.

Figure 7.4 displays the 3-way Venn diagram of these GO term sets. First, we see that most of the GO terms of GLASSO intersect with the terms of the other two methods. Moreover, LRPS and LSCGGM agree on most of their relevant terms with a total of 31 terms in their intersection. There are however 6 terms that are specific

to LSCGGM while 4 terms are found by LRPS but not by LSCGGM. Given the difference in score between these graphs (5.82 vs 4.22), it is interesting to see what these method-specific terms actually are since they must be particularly important. In Table 7.1 (located at the end of this chapter) the GO terms populating some of the subsets of the 3-way Venn diagram are listed. There appear to be three kinds of signal (ordered by decreasing enrichment):

- terms related to bone mineralisation: *e.g.* bone mineralization, calcium-ion binding.
- edges related to transcription and other regulatory mechanisms: *e.g.* RNA processing, mRNA processing, mRNA splicing, RNA splicing, translational initiation.
- some processes involved with viral response: *e.g.* response to virus, defense response to virus, viral transcription, viral process.

There are also a few highly enriched terms that are specific to LSCGGM: *chromatin*, *nucleosomal DNA binding* (related to transcription) and *protein kinase C binding* which is related to bone mineralisation.

Finally, we look at the topology of these graphs. Since there are approximately 200 edges and 800 nodes, many nodes are not connected to anything and plotting the full graphs would not be very informative. In order to select a subset of interesting nodes, we constructed the union graph of the estimates returned by all three methods. We then selected the nodes found in the largest connected component of this union graph: this connected component connects 209 nodes. In Figures 7.5, 7.6 and 7.7 we plot the subgraphs induced by this node set. There is a striking difference between the topology of the graph returned by GLASSO and the ones estimated by the low-rank plus sparse methods: GLASSO identifies hubs with a few nodes connected to many others. This highlights what could be seen as a drawback of low-rank plus

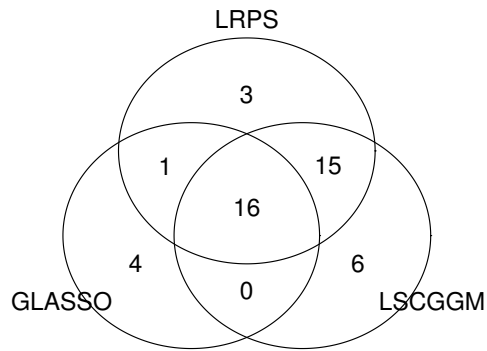


Figure 7.4: Venn diagram of those GO terms that achieved an enrichment statistic of 20 or more in at least one of the three graphs returned by GLASSO, LRPS and LSCGGM. Here, the graphs were selected because they had a number of edges close to 200 (for LRPS and LSCGGM, $\gamma = 0.041$).

sparse approaches: any hub is interpreted as a latent component. It is therefore not surprising to see such a difference between the two families of methods. LRPS and LSCGGM are designed to behave that way.

7.6 Discussion

In light of these results, it is interesting to see the gap in performance between graphical lasso and methods that model latent variables (LRPS, LSCGGM). Indeed, the data used in this analysis was already processed by the authors of Battle et al. (2013) in order to remove hidden variables. To do so they used a method which is considered to be the state of the art when it comes to processing RNA-seq data (Mostafavi et al. 2013). In spite of having regressed out the hidden confounders estimated with their method, we still find that there is some substantial benefit in using low-rank plus sparse approaches. Moreover, this dataset is high-dimensional, which imposes restrictions on the number of components that we can be learned. More remarks pertaining to the results of this chapter and Chapter 6 will be made in the conclusion.

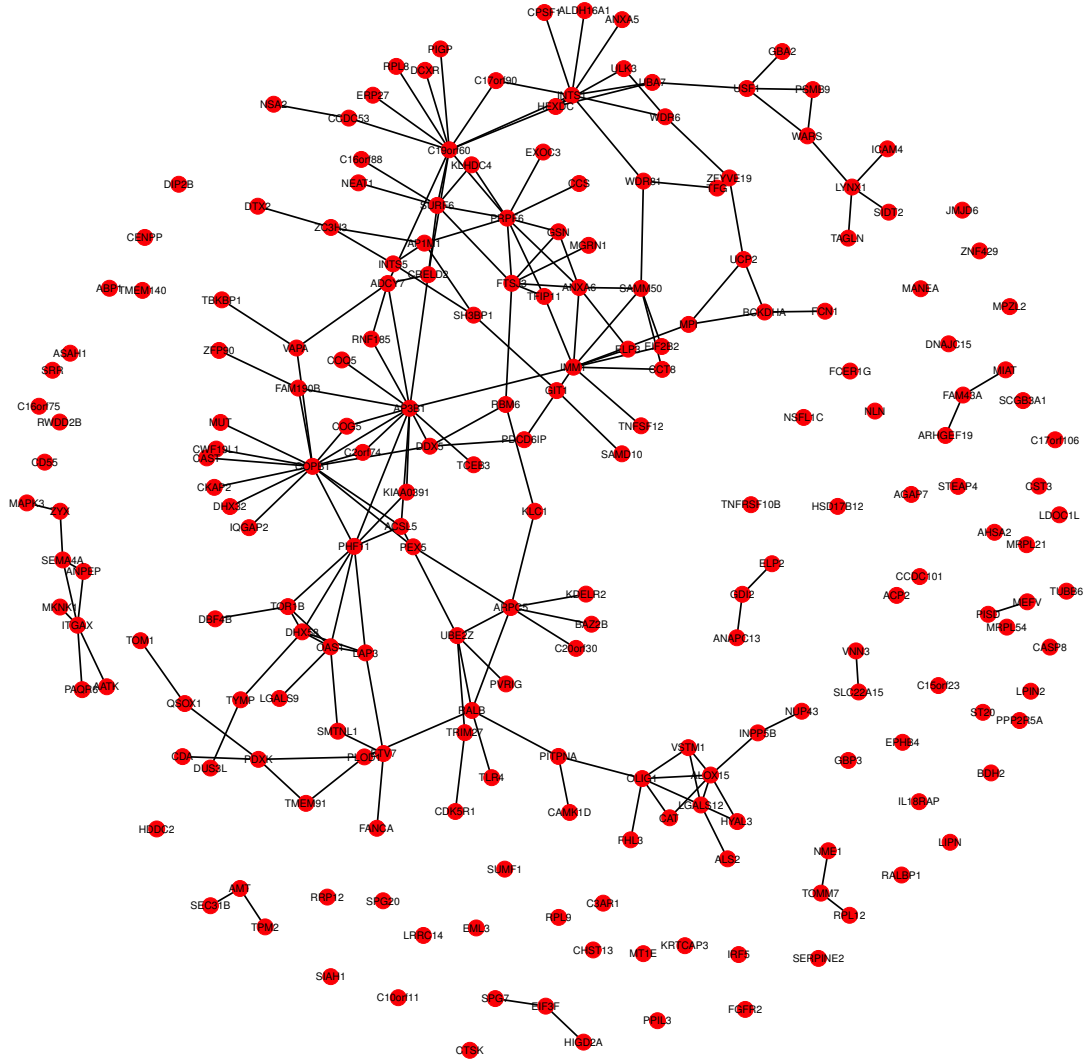


Figure 7.5: Subgraph of the graph estimated by GLASSO. The original graph was selected so that it had approximately 200 edges (204, in this instance). Here we plot the subgraph induced by the largest connected component of the union graph: GLASSO + LRPS + LSCGGM.

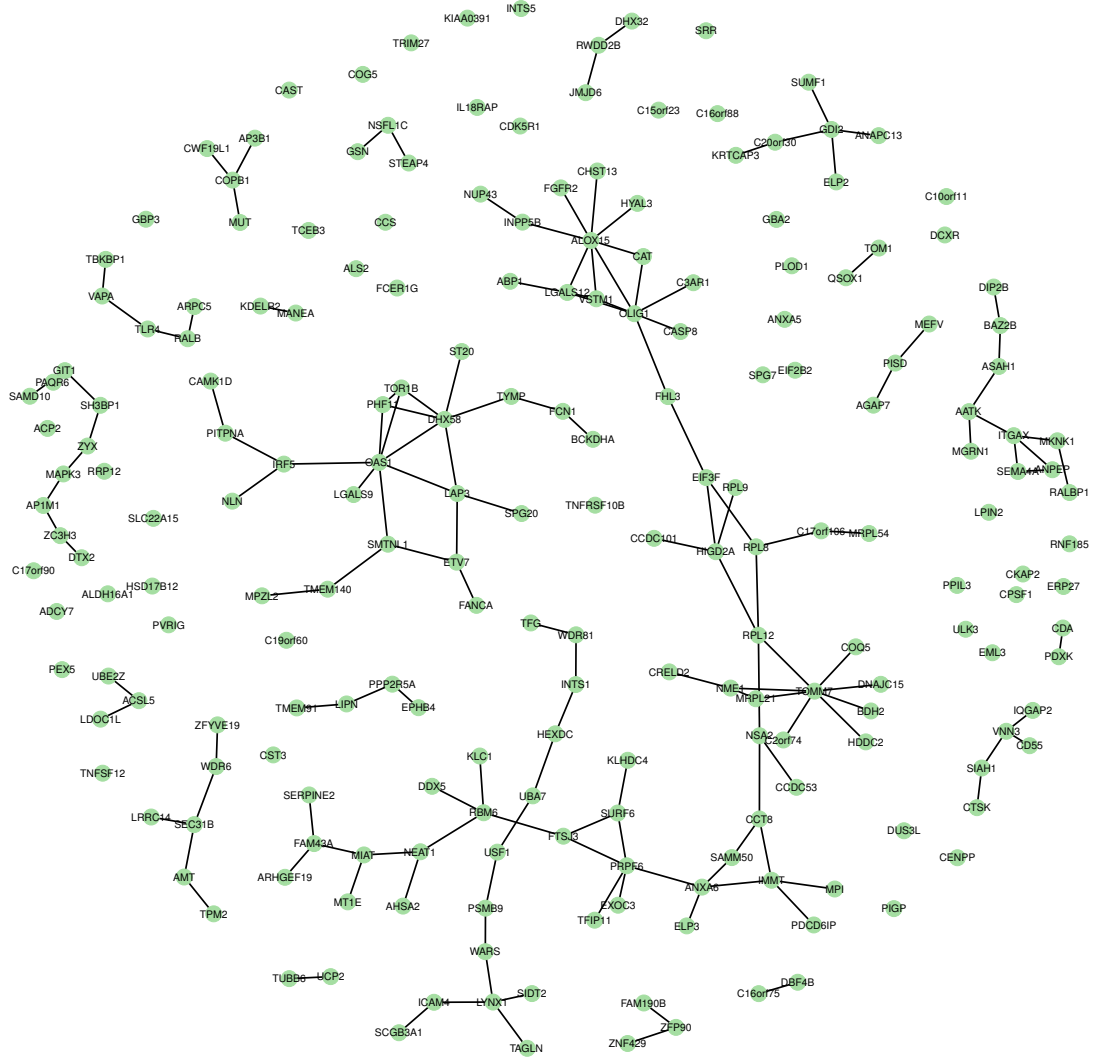


Figure 7.6: Subgraph of the graph estimated by LRPS. The original graph was selected so that it had approximately 200 edges (199, in this instance), with $\gamma = 0.041$. Here we plot the subgraph induced by the largest connected component of the union graph: GLASSO + LRPS + LSCGGM.

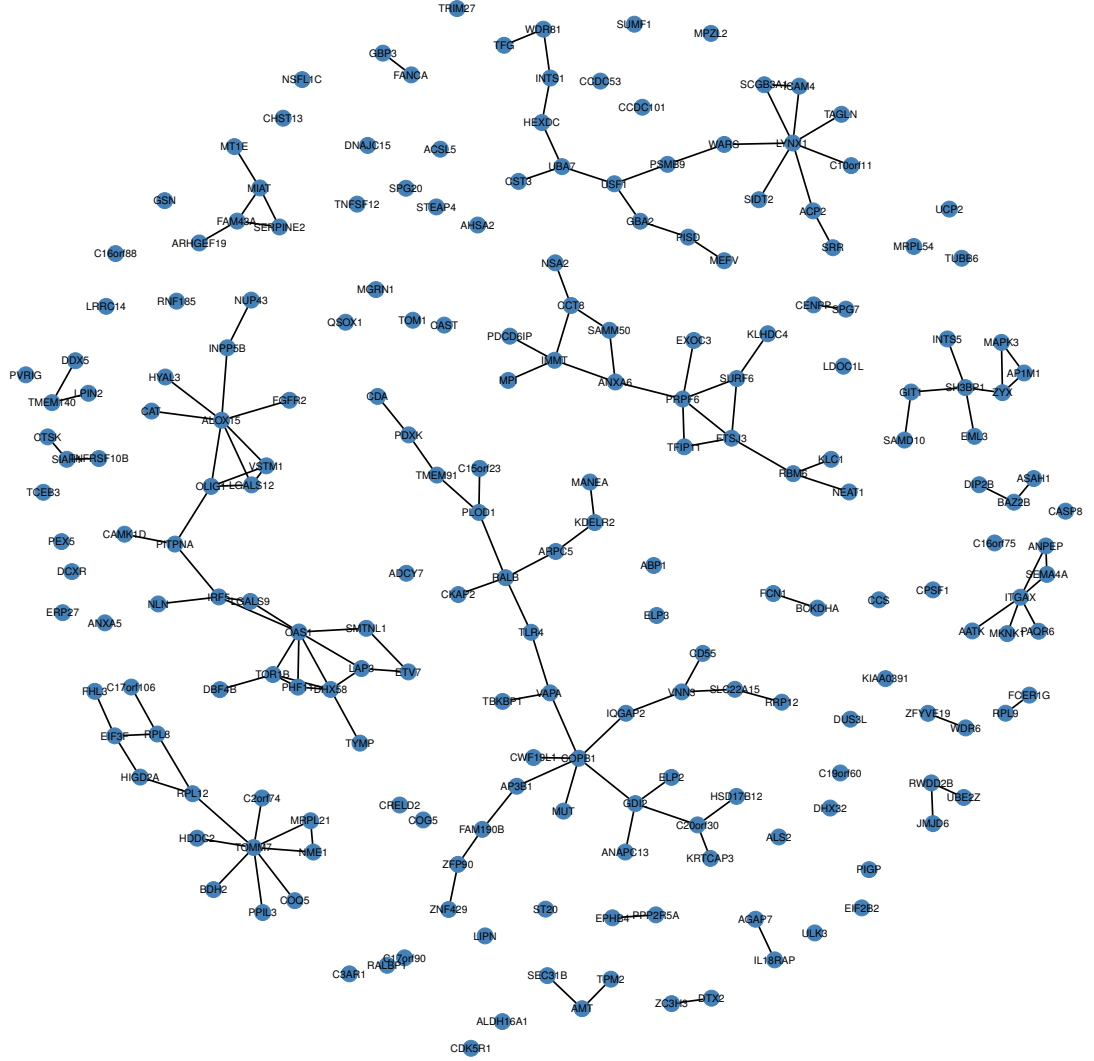


Figure 7.7: Subgraph of the graph estimated by LSCGGM. The original graph was selected so that it had approximately 200 edges (204, in this instance), with $\gamma = 0.041$. Here we plot the subgraph induced by the largest connected component of the union graph: GLASSO + LRPS + LSCGGM.

Table 7.1: GO terms with an enrichment statistic of 20 or more. Graphs were first selected because they had a number of edges close to 200 (for LRPS and LSCGGM, $\gamma = 0.041$). We then consider various subsets of these GO terms by looking at the terms enriched for some graphs and not others. See also Figure 7.4.

Term Set	Term (Ontology)
Found by all three methods	nuclear speck (CC)
	catalytic step 2 spliceosome (CC)
	spliceosomal complex (CC)
	membrane coat (CC)
	GTPase activator activity (MF)
	myelin sheath (CC)
	external side of plasma membrane (CC)
	ribosome biogenesis (BP)
	mitochondrial calcium ion homeostasis (BP)
	defense response to virus (BP)
	response to virus (BP)
	double-stranded RNA binding (MF)
	RNA processing (BP)
	mRNA processing (BP)
	mRNA splicing, via spliceosome (BP)
	RNA splicing (BP)
Found only by LRPS and LSCGGM	rRNA processing (BP)
	nucleolus (CC)
	ribosome (CC)
	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (BP)
	viral transcription (BP)
	interferon-gamma-mediated signaling pathway (BP)
	type I interferon signaling pathway (BP)

	rRNA binding (MF) bone mineralization (BP) viral process (BP) anchored component of membrane (CC) cytosolic large ribosomal subunit (CC) translational initiation (BP) immune response (BP) SRP-dependent cotranslational protein targeting to membrane (BP)
Found only by LRPS	antigen binding (MF) regulation of immune response (BP) structural constituent of ribosome (MF)
Found only by LSCGGM	nucleosomal DNA binding (MF) chromatin (CC) ribonucleoprotein complex (CC) protein kinase C binding (MF) cholesterol metabolic process (BP) dephosphorylation (BP)
Found only by GLASSO	ubiquitin protein ligase binding (MF) cristae formation (BP) calcium-dependent protein binding intra-Golgi vesicle-mediated transport (BP)

Chapter 8

Discussion

We discussed the problem of estimating a conditional Gaussian graphical model in the presence of latent variables. Building on the framework introduced by Chandrasekaran, Parrilo & Willsky (2012), we suggested an estimator which decomposes the parameters of a sparse conditional Gaussian graphical model into the sum of a low-rank and a sparse matrix. Over the course of this thesis, we showed a number of properties of this novel approach:

- Among other theoretical results, we established that the proposed approach is well-behaved in the high-dimensional regime (see Chapter 3).
- We recast our objective function as a semi-definite programming problem. Thanks to this reformulation, published software such as LogDetPPA can be used to fit the model even when the number of variables is in the thousands. In particular, since it uses sparse matrices, it is memory efficient (see Chapter 4).
- Through simulations, we compared the consistency properties of our method to other published approaches. We found that it is capable of correctly recovering the true sparse component of the model in situations in which other approaches cannot. Namely, it can accommodate more latent variables (see Chapter 5).

- When it comes to lasso-type estimators, it is often deplored that appropriate values of the tuning parameters are hard to find. Our simulations also showed that our method is less sensitive to the value of the tuning parameter γ , which makes it easier to set it to a suitable value in real life applications. Moreover, the use of (complementary pairs) stability selection makes the estimates less sensitive to the value of λ while providing some form of error control (see Chapters 4, 5).
- Through two applications to modern datasets comprising genetic and phenotypic measurements, we compared the performances of this approach to alternative methods. We showed how such a conditional graphical model leads to better replication of the results across cohorts and to estimates that are more biologically relevant (see Chapters 6, 7).

The rise of high-throughput genetics, along with progress in data linkage, biobanking and functional genomics projects, has dramatically increased the number of datasets that include both genetic and multivariate phenotypic data. The data application we present in this thesis, using genotype data to draw biological conclusions about the relationships between human traits, is thus becoming one of the most rapidly growing statistical challenges in human genetics. Conditional graphical models are particularly well-suited to such problems as they rely on an assumption we know to be true (namely, that genotype impacts phenotype and not vice-versa). Moreover, genetic measurements are discrete in nature and it is therefore difficult to model them alongside continuous measurements. To the best of our knowledge, there are no approaches capable of learning a joint distribution over continuous and discrete data in the presence of latent variables.

In the introduction, we identified a few of the key challenges of datasets composed of genetic and phenotypic measurements (*e.g.* not all relevant variables are observed, high-dimensionality, etc...) and subsequently decided to focus on two research ques-

tions:

- *Research Question 1: What does the large number of variables being observed enable?*

In other words, what can be done with hundreds of variables that cannot be achieved when modelling the distribution of two variables?

- The direction of effect of some variables is known:

Research Question 2: How can we use our knowledge about the direction of some edges to increase the identifiability of our models?

It is clear that we did not answer these questions in full generality. However, provided we limit ourselves to the class of datasets we are interested in, the estimator suggested here does make a good use of *both* the large number of variables being measured and our knowledge of the problem:

- Ignoring the challenges raised by hidden variables, there is some value in modelling many traits jointly using a graphical model. Indeed, as the number of measurements increases, it is more and more likely that confounders will be observed and their effect accounted for. Automatically identifying the conditional independencies of a large number of variables is therefore useful.
- When there are unobserved variables, decomposing the parameters of the model as a sum of low-rank and sparse matrices is also an approach that becomes more and more useful as the number of variables increases. On the one hand, if the number of latent variables is fixed, increasing the dimensionality makes it easier to infer the footprint of the latent variables on the observed correlation structure. On the other hand, as the number of observed variables increases, the number of latent variables is also allowed to increase, as shown by our theoretical results.

- Finally, by modelling phenotypes conditional on an individual’s DNA, we use our knowledge of the problem, namely that we know something about the direction of some effects (as already stated in the previous paragraph).

While other approaches do have some of these features, the estimator suggested in this thesis is the only one who achieves all three. As was demonstrated in real-world datasets, this results in a better biological relevance of the estimates and better replication across cohorts.

On modelling high-dimensional datasets: screening for putative causal relationships

In Chapters 6 and 7, we compared the graphical lasso and the low-rank plus sparse decomposition to our method in two real-world applications. The main difference between the datasets studied in these two chapters was $\tau = \frac{m+p}{n}$: the ratio between the number of samples (n) and the number of observed variables ($m+p$). In Chapter 6 this ratio was very small (< 0.05) while it was greater than one in Chapter 7. When τ is small, there are enough samples to estimate a complex latent structure and fitting the model for large values of γ is possible (for example, in Chapter 6 $\gamma \simeq 0.95$ did not raise any issues). On the other hand, as τ increases, one finds that the fitting algorithm does not converge when γ is too large: the number of parameters quickly becomes too large for the number of samples. For comparison, in Chapter 7, $\gamma = 0.1$ was already too high a value. This observation highlights one of the limitations of low-rank plus sparse approaches.

To understand what might happen in real world datasets, let us consider a hypothetical dataset which satisfies Assumptions 1 and 2 of Chapter 3: we know that for n large enough, there is a λ^* and a range Γ^* of values of γ such that our method perfectly recovers the graph structure. Let us assume further that we have knowledge of Γ^* . As n gets smaller, we are facing two options:

- We fit the model with a value of $\gamma \in \Gamma^*$: when n gets too small, the estimated sparse component will be empty. By doing so, we are being conservative and acknowledge that we do not have enough data to reliably estimate both the graph and the latent structure. If the data at our disposal was generated by an expensive experiment – *e.g.* the data of Chapter 7 whose cost was around 1.5 million US dollars¹ – it might be frustrating not to be able to claim anything useful about the data (using the suggested approach, at least).
- We deliberately pick a value of γ that we know to be smaller than $\min(\Gamma^*)$ and decrease its value as a function of n , so as to retrieve a reasonable number of edges. Then, we know for a fact that some of our findings will be confounded since the ratio $\frac{\text{number of edges}}{\text{number of latent variables}}$ is higher than it should be. On the other hand, we are still estimating a few latent components and thus accounting for some of the confounding. For that reason, we might still expect our estimates to be enriched in truly causal relationships in comparison to other methods that do not account for hidden variables at all, *e.g.* the graphical lasso.

To these remarks, we add that the conditions required for the consistency of graphical lasso-type estimators are typically not met in real-life datasets (Bühlmann & Van de Geer 2011).

In summary, it is unreasonable to expect perfect recovery in real-life applications. While this is true, a more realistic take is to regard our method as a means to generate “causal” hypotheses from a high-dimensional dataset. Paired with stability selection, such an approach can realistically be used to generate a high-quality set of putative causal relationships that can then further be investigated using hypothesis testing driven approaches (*e.g.* the “Some Invalid Some Valid Instrumental Variables Estimator” mentioned in the introduction (Kang et al. 2016)). As shown in our

¹See https://projectreporter.nih.gov/project_info_description.cfm?aid=7941979&icde=16624082

applications this is an achievable goal.

Goodness of fit and uncertainty quantification

Very little attention has been given to goodness-of-fit measures and uncertainty quantification: we relied exclusively on complementary pairs stability selection in order to select the tuning parameters and provide some form of error control for the estimates. There are, however, multiple ways of assessing the goodness-of-fit of (conditional) Markov random fields. In addition, such measures can often be used to select the tuning parameters λ, γ .

The simplest approach is to compute the Kullback-Leibler divergence between the data and the estimated conditional Gaussian Markov random field. In order to compare the performance of different methods, it is customary to report the Kullback-Leibler divergence of an estimate with the data along with the number of free parameters of the model. Unfortunately, the Kullback-Leibler divergence cannot be computed in the high-dimensional setting. Moreover, such a measure cannot readily or automatically be used to select the value of the tuning parameters.

The Bayesian Information Criterion (BIC) is another popular and computationally efficient approach to assess the goodness-of-fit of a regularised MLE such as the one suggested in this thesis (Gao et al. 2012). Gao et al. (2012) show that when the dimension of the problem p is fixed and $n \rightarrow \infty$, the BIC yields a consistent selection procedure. In the high-dimensional regime, when p is allowed to grow with n , an additional term must be added to the traditional BIC. Examples of definitions and theoretical properties of such *extended BICs* can be found in Chen & Chen (2008), Gao et al. (2012), Foygel & Drton (2010).

Finally, though sometimes computationally burdensome, cross-validation techniques can be used to assess the generalisation power of a method. A common approach is to fit the model to 90% of the data and compute the likelihood of the

estimate in the remaining 10% of the samples. This is particularly effective when one is primarily interested in assessing and/or maximising predictive power.

Estimating Directed Acyclic Graphs

In the introductory chapter, we stated our belief that most biological datasets are generated according to a model similar to the one described in Figure 1.2. Recall that this model was a Directed Acyclic Graph (DAG). We chose to focus on the selection of an undirected graph on the premise that, should the true model be a DAG $\mathcal{G}$, estimating an inverse covariance matrix would provide an estimate of $Mor(\mathcal{G})$, a graph which is related to $\mathcal{G}$. This decision was taken because inverse covariance estimation enjoys computational properties that are better suited to the study of high-dimensional datasets: the objective function is convex and fitting can be performed fast enough to allow the analysis of interesting biological datasets. Most importantly, our method models all variables *jointly* and is thus capable of identifying the correlations induced by hidden variables.

In comparison, one of the most widely used DAG estimation methods is the Greedy Equivalence Search (GES) of Chickering & Boutilier (2002). GES is a score based method which greedily adds edges until a (non-convex) score function (usually the Bayesian Information Criterion (BIC)) cannot be improved by the addition of a new edge: it relies only on *local* information (a node and its parents). Such an approach estimates a *Completed Partially Directed Acyclic Graph* (CPDAG) which encodes the set of all DAGs that are Markov equivalent with the true DAG: some edges are directed while others are left undirected. While interesting, this method cannot cope with latent variables. Moreover, when there are latent variables, the observed precision matrix appears very dense (for example, if the latent variables have a low-rank structure). As a result, GES adds many edges and the algorithm takes a long time before reaching termination. In our experience, this is true even when the

number of variables being modelled is relatively small ($p < 100$, say).

Ongoing work suggests that applying GES to the sparse component estimated by a low-rank plus sparse method yields very good results: the first step identifies long-range correlations and removes them, while GES focuses on learning a DAG conditional of the latent structure learned in the previous stage. Thus, inverse covariance estimation in the presence of latent variables can be seen as a stepping-stone towards DAG estimation. The better the initial estimate of $\hat{S}$ (or $\hat{S}_X$, in the case of our estimator) the better the DAG estimated by GES. For that reason, we believe that the work presented in this thesis can constitute the core of a more sophisticated approach whose output would be a CPDAG. The estimated CPDAG can, in turn, be used to query the causal effect between pairs of variables thanks to the IDA algorithm of Maathuis et al. (2009).

Further work

Naturally, the method suggested here also suffers from a number of limitations and more work is required. For example, assuming that the latent variables are normally distributed appears quite restrictive when compared to the flexibility offered by instrumental variable methods. The question of learning discrete graphical models is also important but it is not yet clear how the present work can be extended to such models.

An interesting direction of research would be to understand when and how a statistical model (not necessarily one that assumes normality) can be decomposed on a basis of low-complexity models. In this particular instance, it was shown by Chandrasekaran, Parrilo & Willsky (2012) that, under some suitable conditions, the varieties of sparse and low-rank matrices intersect transversely, thus making it possible to uniquely decompose the parameter of a Gaussian graphical model on the basis of sparse and low-rank matrices. In a first instance, it might be possible to use a kernel

approach in order to extend such decompositions to other settings. Using elements of information geometry, more general results might also become accessible.

Appendix A

A summary of our notations

This appendix provides a list of the different notation and constants used throughout our proof.

- m : number of variables we condition on. With our notations, $m = |Z|$.
- p : number of variables being modelled. With our notations, $p = |X|$.
- $\|M\|_2 = \sqrt{\text{largest eigenvalue of } M^T M}$: spectral norm of M .
- $\|M\|_\infty$: largest entry in magnitude of M .
- $\|M\|_1 = \sum_{i,j} |M_{ij}|$: sum of the absolute value of the entries of M .
- $\|M\|_* = \sum_i \sigma_i$: sum of the singular values of M .
- $g_\gamma(A, B) = \max(\frac{\|A\|_\infty}{\gamma}, \|B\|_2)$.
- $\mathcal{A}$, the operator that adds two matrices.
- $\mathcal{A}^\dagger$: $\mathcal{A}^\dagger(A) = (A, A)$.
- $\mathcal{P}_T$: orthogonal projector onto the linear subspace T .
- $\rho(M_1, M_2) = \max_{\|N\|_2 \leq 1} \|(\mathcal{P}_{T_1} - \mathcal{P}_{T_2})(N)\|_2$.

- $\mathcal{L}(r)$: variety of matrices of size $(m+p) \times p$ of rank at most r .
- $T(M)$, for M a matrix of rank r : the tangent space to $\mathcal{L}(r)$ at M .
- $\mathcal{S}(k)$: variety of matrices of size $(m+p) \times p$ with at most k non-entries.
- $\Omega(M)$, for M a matrix with k non-entries: tangent space to $\mathcal{S}(k)$ at M .
- $\xi(T(M)) = \max_{N \in T(M), \|N\|_2 \leq 1} \|N\|_\infty$.
- $\mu(\Omega(M)) = \max_{N \in \Omega(M), \|N\|_\infty \leq 1} \|N\|_2$.
- $\mathcal{I}_{\Sigma_Z}^*$: Fisher Information Matrix evaluated at the nominal parameters.
- $\alpha_\Omega \triangleq \min_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_\Omega \mathcal{I}_{\Sigma_Z}^* \mathcal{P}_\Omega(M)\|_\infty$.
- $\delta_\Omega \triangleq \max_{M \in \Omega, \|M\|_\infty = 1} \|\mathcal{P}_{\Omega^\perp} \mathcal{I}_{\Sigma_Z}^* \mathcal{P}_\Omega(M)\|_\infty$.
- $\beta_\Omega \triangleq \max_{M \in \Omega, \|M\|_2 = 1} \|\mathcal{I}_{\Sigma_Z}^*(M)\|_2$.
- $\alpha_T \triangleq \min_{\rho(T, T') < \xi(T)/2} \min_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'} \mathcal{I}_{\Sigma_Z}^* \mathcal{P}_{T'}(M)\|_2$.
- $\delta_T \triangleq \max_{\rho(T, T') < \xi(T)/2} \max_{M \in T', \|M\|_2 = 1} \|\mathcal{P}_{T'^\perp} \mathcal{I}_{\Sigma_Z}^* \mathcal{P}_{T'}(M)\|_2$.
- $\beta_T \triangleq \max_{\rho(T, T') < \xi(T)/2} \max_{M \in T', \|M\|_\infty = 1} \|\mathcal{I}_{\Sigma_Z}^*(M)\|_\infty$.
- $\alpha \triangleq \min(\alpha_\Omega, \alpha_T)$.
- $\beta \triangleq \max(\beta_\Omega, \beta_T)$.
- $\delta \triangleq \max(\delta_\Omega, \delta_T)$.
- ν : a number in $(0, \frac{1}{2}]$ such that $\frac{\delta}{\alpha} \leq 1 - 2\nu$.
- $w = \max(1, \frac{1}{\gamma})$.
- $D = \max(1, \frac{\nu\alpha}{3\beta(2\nu)})$.
- $\psi_Z = \|\Sigma_Z^n\|_2$.

- $\psi_X^* = \|K_X^{*-1}\|_2$.
- $\phi_{ZX}^* = \|K_{ZX}^*\|_2$.
- $\psi = \frac{3}{2}\psi_X^* \sqrt{\left(1 + 2\frac{\psi_Z}{\psi_X^*} \left(1 + \frac{9}{4}\psi_X^* \phi_{ZX}^*\right)^2\right)}$.
- $C_1 = \frac{48}{\alpha} + \frac{1}{\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} \left(1 + \psi_X^* \phi_{ZX}^*\right)^2\right)}$.
- $C_2 = \left(1 + \frac{24(2-\nu)}{\nu}\right) C_1^2 D \psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} \left(1 + \psi_X^* \phi_{ZX}^*\right)^2\right)$.
- $C_3 = C_1 + \frac{3\alpha C_1^2(2-\nu)}{4(3-\nu)}$.
- $\sigma =$ smallest non-zero singular value of L^* .
- $C_4 = \max\{C_2, C_3\}$.
- $C_5 = \frac{C_1 \nu \alpha}{\beta(2-\nu)}$.
- $C_6 = \frac{\alpha \nu}{32(3-\nu)D} \min\left(\frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha \nu}{384D(3-\nu)\psi_X^* \psi^2(1+\frac{\alpha}{6\beta})^2}\right)$.
- σ : smallest singular value of L^* .
- θ : smallest entry in magnitude of S^* .

Appendix B

Proof of Theorem 2

B.1 Goal

Our goal is to provide a proof of Theorem 2. In the first few sections of this appendix, we recall a few important results from the literature and also extend existing results to the case of non-square matrices. The last section builds on all these preliminary results and concludes the proof. The strategy we follow here is similar to the one developed in Chandrasekaran, Parrilo & Willsky (2012), which is itself reminiscent of the approach originally used in Wainwright (2009). However, almost all of our results differ due to the form taken by the likelihood. An overview of this proof technique is given in section B.5.

Throughout, we use the same notations as the ones introduced in Chapter 3.

B.2 Identifiability

As already mentioned our theoretical chapter, the question of identifiability has already been answered in Chandrasekaran, Parrilo & Willsky (2012). The reason for this is that only a number of properties of the Fisher Information Matrix (FIM) appear in the proofs but the actual form of the FIM is irrelevant.

Here, our goal is to recall a proposition of Chandrasekaran, Parrilo & Willsky (2012) which is instrumental in the consistency proof and to which we will refer time and again.

We first introduce a few additional notations. First, let g_γ denote the dual norm of the regularisation function $f_\gamma = \lambda(\gamma \|S\|_1 + \|L\|_*)$:

$$g_\gamma = \max \left(\frac{\|S\|_\infty}{\gamma}, \|L\|_2 \right).$$

We also write $\mathcal{A} : \mathbb{R}^{(m+p) \times p} \times \mathbb{R}^{(m+p) \times p} \rightarrow \mathbb{R}^{(m+p) \times p}$ for the operator that adds two matrices. As usual, $\mathcal{A}^\dagger$ denotes its adjoint : $\mathcal{A}^\dagger : M \mapsto (M, M)$, so that $\mathcal{A}^\dagger \mathcal{A}(A, B) = (A + B, A + B)$.

Then, under Assumptions 1 and 2 we have the following result.

Proposition 1. (*Chandrasekaran, Parrilo & Willsky (2012), Proposition 3.3*)

Let $\mathcal{Y} = \Omega \times T'$ with $\rho(T, T') \leq \frac{\xi(T)}{2}$ (Recall that ρ, α , etc... were defined in Section 3.1). Then

1.

$$\min_{(S, L) \in \mathcal{Y}, \|S\|_\infty = \gamma, \|L\|_2 = 1} g_\gamma(\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A} \mathcal{P}_\mathcal{Y}(S, L)) \geq \frac{\alpha}{2}$$

and specifically, for $(S, L) \in \mathcal{Y}$

$$g_\gamma(\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A} \mathcal{P}_\mathcal{Y}(S, L)) \geq \frac{\alpha}{2} g_\gamma(S, L).$$

2. Writing $\mathcal{Y}^\perp$ for the orthogonal complement of $\mathcal{Y}$, we have

$$\left\| \mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A} \mathcal{P}_\mathcal{Y} (\mathcal{P}_\mathcal{Y} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A} \mathcal{P}_\mathcal{Y})^{-1} \right\|_{g_\gamma \rightarrow g_\gamma} \leq 1 - \nu.$$

And, more specifically,

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(S, L)) \leq (1 - \nu) g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(S, L)).$$

B.3 Elementary properties of the likelihood

We now make a number of straightforward calculations about the likelihood

$$\ell(K_{ZX}, K_X; \Sigma_O^n) = -\log \det K_X + \text{Tr}(\Sigma_X^n K_X + 2K_{ZX}^T \Sigma_Z^n + K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX}).$$

We are interested in computing the gradient and the Hessian of ℓ . In passing, we show that ℓ is *convex*, so that optimisation problem (2.1) – being the sum of three convex functions – is also convex.

Property 1. (*Maximum Likelihood Estimate (M.L.E.)*)

Assuming Σ_Z^* is non-singular and n is large enough, the M.L.E. is well-defined and given by

$$\begin{aligned} \hat{K}_X &= (\Sigma_X^n - \Sigma_{ZX}^n{}^T \Sigma_Z^{n-1} \Sigma_{ZX}^n)^{-1}; \\ \hat{K}_{ZX} &= -\Sigma_Z^{n-1} \Sigma_{ZX}^n \hat{K}_X. \end{aligned}$$

We can now compute the Hessian of ℓ .

Property 2. (*Fisher Information Matrix*)

$$\mathcal{I}_{\Sigma_Z^n}^* = - \begin{pmatrix} K_X^{*-1} \otimes K_X^{*-1} & 0 \\ 0 & 0 \end{pmatrix} - 2K_X^{*-1} \otimes \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^n K_{ZX}^* K_X^{*-1} & -K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^n \\ \cdot & \Sigma_Z^n \end{pmatrix}.$$

Proof. We use differentials.

First, it is easy to see that

$$\begin{aligned} d\ell &= -\text{tr}(K_X^{-1} dK_X) + \text{tr}(\Sigma_X^n dK_X) + 2\text{tr}(\Sigma_{ZX}^n (dK_{ZX})^T) \\ &\quad + 2\text{tr}(\Sigma_Z^n dK_{ZX} K_X^{-1} K_{ZX}^T) - \text{tr}(\Sigma_Z^n K_{ZX} K_X^{-1} dK_X K_X^{-1} K_{ZX}^T). \end{aligned}$$

Likewise, it is straightforward to compute the second order derivatives. To make the computation easier to follow, we break down $d^2\ell$ into its individual components:

$$\begin{aligned}
d(-tr(K_X^{-1}dK_X)) &= tr(K_X^{-1}dK_X A_X^{-1}dK_X); \\
d(tr(\Sigma_X^n dK_X)) &= 0; \\
d(tr(\Sigma_{ZX}^n (dK_{ZX})^T)) &= 0; \\
d(tr(\Sigma_Z^n dK_{ZX} K_X^{-1} K_{ZX}^T)) &= -tr(K_X^{-1}dK_X K_X^{-1} K_{ZX}^T \Sigma_Z^n dK_{ZX}) \\
&\quad + tr(K_X^{-1}dK_{ZX}^T \Sigma_Z^n dK_{ZX}); \\
d(-tr(\Sigma_Z^n K_{ZX} K_X^{-1} dK_X K_X^{-1} K_{ZX}^T)) &= 2tr(K_X^{-1}dK_X K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1} dK_X) \\
&\quad - 2tr(K_X^{-1}dK_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1} dK_X).
\end{aligned}$$

So that:

$$\begin{aligned}
d^2\ell &= tr(K_X^{-1}dK_X K_X^{-1}dK_X) + 2tr(K_X^{-1}dK_{ZX}^T \Sigma_Z^n dK_{ZX}) \\
&\quad + 2tr(K_X^{-1}dK_X K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1} dK_X) \\
&\quad - 4tr(K_X^{-1}dK_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1} dK_X).
\end{aligned}$$

Now, write $K := \begin{pmatrix} K_X \\ K_{ZX} \end{pmatrix}$ and similarly $vec(K) := \begin{pmatrix} vec(K_X) \\ vec(K_{ZX}) \end{pmatrix}$. Then, using the identity $tr(ABCD) = (vec B^T)^T (A^T \otimes C) vec D$, we obtain

$$\begin{aligned}
tr(K_X^{-1}dK_X K_X^{-1}dK_X) &= d(vec K_X)^T (K_X^{-1} \otimes K_X^{-1}) dvec K_X \\
tr(K_X^{-1}dK_{ZX}^T \Sigma_Z^n dK_{ZX}) &= d(vec K_{ZX})^T (K_X^{-1} \otimes \Sigma_Z^n) dvec K_{ZX} \\
tr(K_X^{-1}dK_X K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1} dK_X) &= d(vec K_X)^T (K_X^{-1} \otimes K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1}) dvec K_X \\
tr(K_X^{-1}dK_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1} dK_X) &= d(vec K_{ZX})^T (K_X^{-1} \otimes \Sigma_Z^n K_{ZX} K_X^{-1}) dvec K_X.
\end{aligned}$$

So that,

$$\begin{aligned}
d^2\ell &= d(\text{vec}K_X)^T \left[(K_X^{-1} \otimes K_X^{-1}) + 2(K_X^{-1} \otimes K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1}) \right] d\text{vec}K_X \\
&\quad 2d(\text{vec}K_{ZX})^T (K_X^{-1} \otimes \Sigma_Z^n) d\text{vec}K_{ZX} \\
&\quad -4d(\text{vec}K_{ZX})^T (K_X^{-1} \otimes \Sigma_Z^n K_{ZX} K_X^{-1}) d\text{vec}K_X
\end{aligned}$$

and

$$d^2\ell = (d\text{vec}K)^T \begin{pmatrix} K_X^{-1} \otimes K_X^{-1} + 2K_X^{-1} \otimes K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1} & -2K_X^{-1} \otimes \Sigma_Z^n K_{ZX} K_X^{-1} \\ \cdot & 2K_X^{-1} \otimes \Sigma_Z^n \end{pmatrix} d\text{vec}K.$$

From which the result follows. $\square$

In passing, remark that $-\mathcal{I}_{\Sigma_Z}(K)$ is positive semi-definite if and only if

$$\begin{pmatrix} K_X^{-1} K_{ZX}^T \Sigma_Z^n K_{ZX} K_X^{-1} & -K_X^{-1} K_{ZX}^T \Sigma_Z^n \\ \cdot & \Sigma_Z^n \end{pmatrix}$$

is positive semi-definite. This is because K_X is positive definite (we have the constraint $K_X \succ 0$). But it is easy to see that this matrix is simply a covariance matrix of the form $A^T A$. In conclusion, ℓ is convex and so is (2.1).

B.4 Curvature of the likelihood and the rank variety

B.4.1 Curvature of the rank variety

The goal of this section is to extend two results of Chandrasekaran, Parrilo & Willsky (2012) from the special case of positive definite matrices to arbitrary matrices. We rely on previous work by Bach (2008).

We give here the results that we seek to prove and dedicate the rest of this section to their proof.

Given two linear subspaces T_1, T_2 of same dimension define

$$\rho(T_1, T_2) \triangleq \|\mathcal{P}_{T_1} - \mathcal{P}_{T_2}\|_{2 \rightarrow 2} = \max_{\|N\|_2 \leq 1} \|(\mathcal{P}_{T_1} - \mathcal{P}_{T_2})(N)\|_2,$$

which measures the “angle” between these two subspaces. Recall that for any matrix $M \in \mathbb{R}^{p \times q}$, $T(M)$ denotes the tangent space to the variety of low-rank matrices at M . Then we have the following proposition.

Proposition 2. (*Extension of Chandrasekaran, Parrilo & Willsky (2012) (suppl. mat.), Proposition 2.1, 2.2.*)

Let $W \in \mathbb{R}^{p \times q}$ be a rank $r < \min(p, q)$ matrix with non-zero singular values $(\sigma_i)_{i \in \{1, \dots, r\}}$. Write $\sigma = \min_i \sigma_i$, and let Δ be such that $\|\Delta\|_2 < \frac{\sigma}{4}$. Let $W + \Delta$ be a rank r matrix. Then we have

1.

$$\rho(T(W + \Delta), T(W)) \leq \frac{8}{\sigma} \|\Delta\|_2;$$

2.

$$\|\mathcal{P}_{T(W)^\perp}(\Delta)\|_2 \leq \frac{4}{\sigma} \|\Delta\|_2^2.$$

B.4.1.1 Jordan-Wieland Matrix and Matrix perturbation bounds:

Throughout, we assume that W is some $\mathbb{R}^{p \times q}$ matrix, with non-zero singular values $\sigma_i, i = 1, \dots, r$ (indexed by decreasing order) and singular vectors, u_i, v_i . The corresponding Jordan-Wiedland matrix is defined from W as follows (Stewart & Sun 1990):

$$\bar{W} \begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix}.$$

This matrix is interesting because it has eigenvalues $\sigma_i, -\sigma_i$ and its eigenvectors are $\frac{1}{\sqrt{2}} \begin{pmatrix} u_i \\ \pm v_i \end{pmatrix}$. We write $\bar{W} = \bar{U} \bar{S} \bar{U}^T$ for the eigenvalue decomposition of $\bar{W}$. If the

singular value decomposition of W is $W = USV^T$, then we have $\bar{S} = \frac{1}{\sqrt{2}} \begin{pmatrix} S & 0 \\ 0 & -S \end{pmatrix}$

and $\bar{S} = \frac{1}{\sqrt{2}} \begin{pmatrix} U & U \\ V & -V \end{pmatrix}$ so that $\bar{U} \bar{U}^T = \begin{pmatrix} UU^T & 0 \\ 0 & VV^T \end{pmatrix}$.

Now, let Δ be a small perturbation to W . If $\|\Delta\|_2 \leq \frac{\sigma_r}{2}$ then $W + \Delta$ has r singular values that are strictly greater than $\sigma_r/2$ and all the remaining ones are strictly less than $\sigma_r/2$ (Stewart & Sun 1990). Let $P_{\bar{W}, \sigma_r/2}$ denote the projector on the $p + q - 2r$ -dimensional invariant subspace of $\bar{W}$ which corresponds to the smallest eigenvalues (in this case it is a null space as W is of rank exactly r). Likewise, if $\|\Delta\|_2 \leq \sigma_r/2$, $P_{\bar{W} + \bar{\Delta}, \sigma_r/2}$ also denotes a projector onto a $p + q - 2r$ dimensional subspace. The projection onto the orthogonal subspace is given by $I - P_{\bar{W} + \bar{\Delta}, \sigma_r/2}$.

Bach (2008) shows the following results

Proposition 3. (*Bach (2008), Proposition 16*)

Assume W is of rank r and $\|\Delta\|_2 < \frac{\sigma_r}{4}$. Then, the projection on the first r eigenvectors of $\bar{W}$, $I - P_{\bar{W}, \sigma_r/2}$, is such that

$$\|P_{\bar{W} + \bar{\Delta}, \sigma_r/2} - P_{\bar{W}, \sigma_r/2}\|_2 \leq \frac{4}{\sigma_r} \|\Delta\|_2.$$

Now, recall that for any two matrices A, B (of compatible dimensions) we have the following inequalities:

$$\max(\|A\|_2, \|B\|_2) \leq \left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\|_2 \leq \|A\|_2 + \|B\|_2.$$

Writing $P_{U(W)}$ (resp. $P_{V(W)}$) for the projection onto the row-space $U(W)$ (resp.

$(V(M)))$ then we have the following corollary

Corollary 1. *Assume W is of rank r and $\|\Delta\|_2 < \frac{\sigma_r}{4}$. Assume further that $\text{rank}(W + \Delta) = r$. Then we have that*

$$\max(\|P_{U(W+\Delta)} - P_{U(W)}\|_2, \|P_{V(W+\Delta)} - P_{V(W)}\|_2) \leq \|P_{\bar{W}+\bar{\Delta}, \sigma_r/2} - P_{\bar{W}, \sigma_r/2}\|_2.$$

We also have the following result:

Proposition 4. *(Bach (2008), Proposition 18)*

Assume that W has rank $r < \min(p, q)$, with singular value decomposition $W = USV^T$. If $\frac{4}{\sigma_r} \|\Delta\|_2^2 < \|(I - UU^T)\Delta(I - VV^T)\|_2$, then $\text{rank}(W + \Delta) > r$.

B.4.1.2 Curvature of the rank variety – Proof of Proposition 2

Let W be a $p \times q$ matrix defined as before. The projection onto the matrix variety of low-rank matrices at W , $T(W)$, is written $\mathcal{P}_{T(W)}$. For any matrix N , it can be expressed using the row/column projectors as follows (Chandrasekaran et al. (2009)):

$$\mathcal{P}_{T(W)}(N) = P_{U(W)}N + NP_{V(W)} - P_{U(W)}NP_{V(W)},$$

while the projector onto the orthogonal subspace is $(I - \mathcal{P}_{T(W)})$, i.e.:

$$\mathcal{P}_{T(W)^\perp} = (I - P_{U(W)})N(I - P_{V(W)}).$$

For any matrix N we have that

$$\begin{aligned} (\mathcal{P}_{T(W+\Delta)} - \mathcal{P}_{T(W)}) = \\ (P_{U(W+\Delta)} - P_{U(W)})N(I - P_{V(W)}) + (I - P_{U(W+\Delta)})N(P_{V(W+\Delta)} - P_{V(W)}). \end{aligned}$$

As a result, under the assumptions of Proposition 2 (in particular that $\|\Delta\|_2 \leq$

$\sigma_r/4$), we have the following inequalities:

$$\begin{aligned}
\rho(T(W + \Delta), T(W)) &\leq \max_{\|N\|_2 \leq 1} \left\| (P_{U(W+\Delta)} - P_{U(W)}) N (I - P_{V(W)}) + \right. \\
&\quad \left. (I - P_{U(W+\Delta)}) N (P_{V(W+\Delta)} - P_{V(W)}) \right\|_2 \\
&\leq \max_{\|N\|_2 \leq 1} \left\| (P_{U(W+\Delta)} - P_{U(W)}) N (I - P_{V(W)}) \right\|_2 + \\
&\quad \max_{\|N\|_2 \leq 1} \left\| (I - P_{U(W+\Delta)}) N (P_{V(W+\Delta)} - P_{V(W)}) \right\|_2 \\
&\leq 2 \max(\|P_{U(W+\Delta)} - P_{U(W)}\|, \|P_{V(W+\Delta)} - P_{V(W)}\|) \\
&\leq \frac{8}{\sigma_r} \|\Delta\|_2,
\end{aligned}$$

where we used the corollary given earlier. This proves part 1) of Proposition 2.

We now turn to part 2), where we wish to prove that

$$\|\mathcal{P}_{T(W)^\perp}(\Delta)\|_2 \leq \frac{4}{\sigma} \|\Delta\|_2^2.$$

This is a direct consequence of Proposition 3. Indeed, we have that

$$\|\mathcal{P}_{T(W)^\perp}(\Delta)\|_2 = \|(I - P_{U(W)})\Delta(I - P_{V(W)})\|_2,$$

which – using the contra-positive of Proposition 3 – concludes the proof, since we assumed that $W + \Delta$ was a rank r matrix.

B.4.1.3 Curvature of the likelihood

Throughout the rest of the appendix we will make use of the following notations.

They have all been defined in Chapter 3.

Let $w = \max(1, \frac{1}{\gamma})$ and set $D = \max(1, \frac{\nu\alpha}{3\beta(2-\nu)})$. We assume that

$$\gamma \in \left[\frac{3\xi(T)\beta(2-\nu)}{\nu\alpha}, \frac{\nu\alpha}{2\mu(\Omega)\beta(2-\nu)} \right],$$

so that $w \leq \frac{D}{\xi(T)}$.

In order to bound the error terms, it is necessary to study the curvature of the likelihood gradient at the true parameters. In particular, we want to know how “close” (here taken in the $\|\cdot\|_2$ sense) to the true parameters the estimates have to be in order to achieve a particular error bound.

To that end we introduce the following matrix valued function:

$$\mathcal{F}_{\Sigma_Z^n} : \mathbb{R}^{(m+p) \times p} \rightarrow \mathbb{R}^{(m+p) \times p}$$

$$M = \begin{pmatrix} M_X \\ M_{ZX} \end{pmatrix} \mapsto \begin{pmatrix} \mathcal{F}_{\Sigma_Z^n, X}(M) \\ \mathcal{F}_{\Sigma_Z^n, ZX}(M) \end{pmatrix} \triangleq \begin{pmatrix} M_X^{-1} + M_X^{-1} M_{ZX}^T \Sigma_Z^n M_{ZX} M_X^{-1} \\ -2 \Sigma_Z^n M_{ZX} M_X^{-1} \end{pmatrix}.$$

We can now express the following proposition which will be used later in order to bound the error terms.

Proposition 5. *Consider the Taylor expansion of $\mathcal{F}_{\Sigma_Z^n}$ at $K^* = \begin{pmatrix} K_X^* \\ K_{ZX}^* \end{pmatrix}$:*

$$\mathcal{F}_{\Sigma_Z^n}(K^* + \Delta) = \mathcal{F}_{\Sigma_Z^n} K^* + d(\mathcal{F}_{\Sigma_Z^n} K^*) \Delta + R_{K^*}(\Delta),$$

(note that in order to avoid cluttered notations we do not write $R_{K^*, \Sigma_Z^n}(\Delta)$, but the reader should not forget that there is a dependency here.). Write $\psi_Z \triangleq \|\Sigma_Z^n\|_2$, $\psi_X^* \triangleq \|K_X^{*-1}\|_2$ and $\phi_{ZX}^* \triangleq \|K_{ZX}^*\|_2$. If $\|\Delta\|_2 \leq \min(\frac{1}{3\psi_X^*}, \frac{\phi_{ZX}^*}{2})$ then

$$\|R_{K^*}(\Delta)\|_2 \leq 3\psi_X^* \psi^2 \|\Delta\|_2^2,$$

where

$$\psi \triangleq \frac{3}{2} \psi_X^* \sqrt{\left(1 + 2 \frac{\psi_Z}{\psi_X^*} \left(1 + \frac{9}{4} \psi_X^* \phi_{ZX}^*\right)^2\right)}.$$

Proof. We start by recalling that for any two matrices A, B (of compatible dimen-

sions) we have the following inequalities:

$$\max(\|A\|_2, \|B\|_2) \leq \|C\|_2 \leq \|A\|_2 + \|B\|_2,$$

$$\text{where } C = \begin{pmatrix} A \\ B \end{pmatrix}.$$

By the mean value theorem, there exists a real t , $0 \leq t \leq 1$, such that $R_{K^*}(\Delta) = d^2(\mathcal{F}_{\Sigma_Z^n}(K^* + t\Delta); \Delta)$, which is the second derivative at $K^* + t\Delta$ evaluated at Δ . Given the inequality above, it is clear that, in order to obtain a bound on the norm of the remainder, it is enough to bound the sum $\|R_{X,K^*}(\Delta_X)\|_2 + \|R_{ZX,K^*}(\Delta_{ZX})\|_2$. This is achieved by bounding both terms individually.

A tedious calculation (similar in all points to the one given in the appendix of Wytock & Kolter (2013)) yields the following expressions:

$$\begin{aligned} d^2(\mathcal{F}_{X,\Sigma_Z^n}(M); \Delta) &= 2(M_X^{-1}\Delta_X M_X^{-1}\Delta_X M_X^{-1} + M_X^{-1}\Delta_X M_X^{-1}\Delta_X M_X^{-1}M_{ZX}^T \Sigma_Z^n M_{ZX} M_X^{-1} + \\ &\quad M_X^{-1}\Delta_X M_X^{-1}M_{ZX}^T \Sigma_Z^n M_{ZX} M_X^{-1}\Delta_X M_X^{-1} + \\ &\quad M_X^{-1}M_{ZX}^T \Sigma_Z^n M_{ZX} M_X^{-1}\Delta_X M_X^{-1}\Delta_X M_X^{-1} - \\ &\quad M_X^{-1}\Delta_X M_X^{-1}\Delta_{ZX}^T \Sigma_Z^n M_{ZX} M_X^{-1} - M_X^{-1}\Delta_X M_X^{-1}M_{ZX}^T \Sigma_Z^n \Delta_{ZX} M_X^{-1} - \\ &\quad M_X^{-1}\Delta_{ZX} \Sigma_Z^n M_{ZX} \Delta_X M_X^{-1} - M_X^{-1}M_{ZX}^T \Sigma_Z^n \Delta_{ZX} M_X^{-1}\Delta_X M_X^{-1} + \\ &\quad M_X^{-1}\Delta_{ZX}^T \Sigma_Z^n \Delta_{ZX} M_X^{-1}). \\ d^2(\mathcal{F}_{ZX,\Sigma_Z^n}(M); \Delta) &= 2(-2\Sigma_Z^n M_{ZX} M_X^{-1}\Delta_X M_X^{-1}\Delta_X M_X^{-1} + 2\Sigma_Z^n \Delta_{ZX} M_X^{-1}\Delta_X M_X^{-1}) \end{aligned}$$

We are interested in bounding $\|d^2(\mathcal{F}_{\Sigma_Z^n}(K^* + t\Delta); \Delta)\|_2$ and the terms $\|(K_X^* + t\Delta_X)^{-1}\|_2$, $\|K_{ZX}^* + t\Delta_{ZX}\|_2$ appear many times in these expressions. We use the assumption on $\|\Delta\|_2$ to show that $R_{K^*,\Sigma_Z^n}(\Delta)$ converges and to bound these two terms.

- Rewrite $(K_X^* + t\Delta_X)^{-1}$ as $K_X^{*-1}(I + tK_X^{*-1}\Delta_X)^{-1}$. Using the submultiplicative property of the spectral norm and the fact that $(I + tK_X^{*-1}\Delta_X)^{-1} =$

$\sum_{i=0}^{\infty} (-1)^i (tK_X^{*-1}\Delta_X)^i$, we obtain :

$$\|(K_X^* + t\Delta_X)^{-1}\|_2 \leq \psi_X^* \frac{1}{1 - \psi_X^* \|\Delta_X\|_2};$$

which, by our assumptions on $\|\Delta\|_2$, implies

$$\|(K_X^* + t\Delta_X)^{-1}\|_2 \leq \frac{3}{2} \psi_X^*.$$

- On the other hand, we have

$$\|K_{ZX}^* + t\Delta_{ZX}\|_2 \leq \|K_{ZX}^*\|_2 + \|\Delta_{ZX}\|_2 \leq \frac{3}{2} \phi_{ZX}^*.$$

Using these two inequalities, it is now straightforward to bound the remainder by bounding its summands independently. Putting together similar terms and rewriting the expression, we have

$$\|R_{K^*}(\Delta)\|_2 \leq 2\left(\frac{3}{2}\psi_X^*\right)^3 \left(1 + 3\frac{\psi_Z}{\frac{3}{2}\psi_X^*} \left(1 + \left(\frac{3}{2}\right)^2 \psi_X^* \phi_{ZX}^*\right)^2\right) \max(\|\Delta_X\|_2^2, \|\Delta_{ZX}\|_2^2),$$

which completes the proof. $\square$

As a corollary, we can prove a result which is similar to Proposition 3.1 of Chandrasekaran, Parrilo & Willsky (2012), (suppl. mat.).

Corollary 2. *Suppose that γ is in the range required for identifiability. Let $g_\gamma(\Delta_S, \Delta_L) \leq \min(\frac{1}{3\psi_X^*}, \frac{\phi_{ZX}^*}{2}) \frac{1}{1 + \frac{\alpha}{6\beta}}$, for any (Δ_S, Δ_L) with $\Delta_S \in \Omega$. Then we have*

$$g_\gamma(\mathcal{A}^\dagger R_{K^*} \mathcal{A}(\Delta_S, \Delta_L)) \leq 3 \frac{D}{\xi(T)} \psi_X^* \psi^2 \left(1 + \frac{\alpha}{6\beta}\right)^2 g_\gamma(\Delta_S, \Delta_L)^2.$$

Proof. Following the proof given in Chandrasekaran, Parrilo & Willsky (2012), we

derive a result that relates the dual norm g_γ to the spectral norm :

$$\begin{aligned}
\|\mathcal{A}(\Delta_S, \Delta_L)\|_2 &\leq \|\Delta_S\|_2 + \|\Delta_L\|_2 \\
&\leq \gamma\mu(\Omega)\frac{\|\Delta_S\|_\infty}{\gamma} + \|\Delta_L\|_2 \\
&\leq (1 + \gamma\mu(\Omega))g_\gamma(\Delta_S, \Delta_L) \\
&\leq (1 + \frac{\alpha}{6\beta})g_\gamma(\Delta_S, \Delta_L) \\
&\leq \min(\frac{1}{3\psi_X^*}, \frac{\phi_{ZX}^*}{2}),
\end{aligned}$$

where we used the range of γ and the assumption on g_γ . Therefore, the assumptions of Proposition 5 are met and we have:

$$\|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2 \leq 3\psi_X^*\psi^2\|\mathcal{A}(\Delta_S, \Delta_L)\|_2^2$$

which implies

$$\|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2 \leq 3\psi_X^*\psi^2(1 + \frac{\alpha}{6\beta})^2g_\gamma(\Delta_S, \Delta_L)^2.$$

Finally, from the definition of $\mathcal{A}^\dagger$ of g_γ , we have

$$g_\gamma(\mathcal{A}^\dagger R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))) = \max(\frac{1}{\gamma}\|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_\infty, \|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2).$$

Recall also the definition of $w = \max(1/\gamma, 1)$ and the fact that $w \leq \frac{D}{\xi(T)}$. Then,

$$\max(\frac{1}{\gamma}\|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_\infty, \|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2) \leq \frac{D}{\xi(T)}\|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2,$$

and the result follows from the bound on $\|R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2$. $\square$

B.5 Consistency

This section studies a variant of estimator (2.1) in which the constraint $L_X \succeq 0$ is removed:

$$\begin{aligned}
 (\hat{S}_X, \hat{L}_X, \hat{S}_{ZX}, \hat{L}_{ZX}) = \\
 \arg \min_{S_X, L_X, S_{ZX}, L_{ZX}} \ell(S_{ZX} - L_{ZX}, S_X - L_X; \Sigma_O^n) + \lambda(\gamma \|S\|_1 + \|L\|_*) \\
 \text{s.t } S_X - L_X \succ 0 \text{ and } S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}. \quad (\text{B.1})
 \end{aligned}$$

Our proof of consistency follows the same pattern as the one in Chandrasekaran, Parrilo & Willsky (2012). Because the likelihood is different and because we consider rectangular matrices (the upper part of which is positive definite), the actual statements of the theorems and propositions are almost always different. In some cases, however, the proof is easily adapted from Chandrasekaran, Parrilo & Willsky (2012) without any major difficulty. For that reason, we will focus on points that are critical to this particular analysis.

Compared to the standard approach to prove the consistency of lasso-type estimators, such as the one used in Ravikumar et al. (2011), Wytock & Kolter (2013), this proof strategy is more involved. Indeed, since the variety of sparse matrices has zero-curvature in its smooth points, tangent space constraints and variety constraints are equivalent. Unfortunately, we also have to deal with the rank variety which has non-zero curvature, hence the need to control for all the tangent spaces that are close to the nominal T .

A detailed description of the proof strategy, along with the rationale behind this approach, is given in Chandrasekaran, Parrilo & Willsky (2012). The key steps are the following:

- We start by considering a version of (B.1) in which the tangent space constraints are enforced explicitly. In its resort to Brouwer's fixed point theorem, this part is fairly similar to the proof given in Ravikumar et al. (2011), Wytock & Kolter (2013).
- We then consider a problem in which the variety constraints are explicitly enforced and show that the optimum of that non-convex problem is algebraically consistent.
- We then show that the optima of the variety constrained and tangent space constrained problems are identical.
- Finally, we derive conditions for the optimum of (B.1) to be identical to the optimum of the tangent space constrained problem.

B.5.1 Statement of the result

First, let us clearly establish the result we seek to prove. We are simply recalling the theorem and the constants introduced in Chapter 3.

We define the following quantities : $D = \max(1, \frac{\nu\alpha}{3\beta(2-\nu)})$, $\psi_Z = \|\Sigma_Z\|_2$, $\psi_X^* = \|K_X^{*-1}\|_2$ and $\phi_{ZX}^* = \|K_{ZX}^*\|_2$, $\psi = \frac{3}{2}\psi_X^* \sqrt{\left(1 + 2\frac{\psi_Z}{\psi_X^*} \left(1 + \frac{9}{4}\psi_X^* \phi_{ZX}^*\right)^2\right)}$. We also set,

$$\begin{aligned}
C_1 &= \frac{48}{\alpha} + \frac{1}{\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right)}, \\
C_2 &= \left(1 + \frac{24(2-\nu)}{\nu}\right) C_1^2 D \psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right), \\
C_3 &= C_1 + \frac{3\alpha C_1^2 (2-\nu)}{4(3-\nu)}, C_4 = \max\{C_2, C_3\}, C_5 = \frac{C_1 \nu \alpha}{\beta(2-\nu)}, \\
C_6 &= \frac{\alpha \nu}{32(3-\nu)D} \min \left(\frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha \nu}{384D(3-\nu)\psi_X^* \psi^2 (1 + \frac{\alpha}{6\beta})^2} \right).
\end{aligned}$$

Finally, set:

$$\delta_n = \sqrt{\frac{256\psi_X^{*2}pM}{n}},$$

with $M = \max\left(1, \frac{\psi_Z}{4\psi_X^*}(1 + \sqrt{\frac{m}{p}})^2\right)$. And, let

$$\lambda_n = \frac{6D(2 - \nu)\delta_n}{\xi(T)\nu}.$$

We want to show the following:

Theorem 5. (*Algebraic Consistency*)

Suppose that Assumptions 1 and 2 hold and that we are given n samples drawn according to

$$Y_X|Y_Z \sim \mathcal{N}\left(-(S_X^* - L_X^*)^{-1}(S_{ZX}^* - L_{ZX}^*)^T Y_Z, (S_X^* - L_X^*)^{-1}\right).$$

Further assume that the following hold:

1. $n \geq \frac{pM}{\xi(T)^4} \max\left(2, \frac{256\psi_X^{*2}}{C_6^2}\right)$.
2. Let the minimum non-zero singular value of L^* be such that

$$\sigma \geq \frac{C_4\lambda_n}{\xi(T)^2}.$$

3. Let the minimum magnitude nonzero entry θ of S^* be such that

$$\theta \geq \frac{C_5\lambda_n}{\mu(\Omega)}.$$

Then, with probability greater than $1 - 2 \min\left(\exp(-pM), \exp\left(-\frac{4\psi_X^*}{\psi_Z}pM\right)\right)$, we have that

1. $\text{sign}(\hat{S}) = \text{sign}(S^*)$ and $\text{rank}(\hat{L}) = \text{rank}(L^*)$

2.

$$\max \left(\frac{1}{\gamma} \left\| \hat{S} - S^* \right\|_{\infty}, \left\| \hat{L} - L^* \right\|_2 \right) \leq \frac{32(3 - \nu)}{3\alpha(2 - \nu)} \lambda_n.$$

In particular, due to the form taken by δ_n ,

$$\max \left(\frac{1}{\gamma} \left\| \hat{S} - S^* \right\|_{\infty}, \left\| \hat{L} - L^* \right\|_2 \right) \leq C \sqrt{\frac{pM}{n}},$$

for some constant C .

B.5.2 Tangent space constraints

We consider

$$\begin{aligned} (\hat{S}_{\Omega}, \hat{L}_{T'}) &= \arg \min_{S, L} -\log \det(S_X - L_X) + \text{tr}(\Sigma_X^n(S_X - L_X)) + 2\text{tr}(\Sigma_{ZX}^n(S_{ZX} + L_{ZX})^T) \\ &\quad + \text{tr}(\Sigma_Z^n(S_{ZX} + L_{ZX})(S_X - L_X)^{-1}(S_{ZX} + L_{ZX})^T) \\ &\text{s.t. } S_X - L_X \succ 0, S \in \Omega, L \in T' \\ &\text{where } S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}; \end{aligned} \tag{B.2}$$

for some T' . The goal of this section is to show that if T' is sufficiently close to the nominal T (as measured by $\rho(T, T')$), then the error terms are bounded.

Let $C_{T'} = \mathcal{P}_{T'}(L^*)$ be the orthogonal projection of the nominal low-rank matrix onto the linear subspace T' .

Proposition 6. *Let the errors (Δ_S, Δ_L) be defined as above and assume that T' is such that $\rho(T, T') \leq \frac{\xi(T)}{2}$. Define*

$$E_n = \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n} \left(\begin{pmatrix} K_X^* \\ K_{ZX}^* \end{pmatrix} \right)$$

and

$$r = \max \left(\frac{8}{\alpha} \left(g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{C}_{T'}) + \lambda_n \right), \|\mathcal{C}_{T'}\|_2 \right).$$

If

$$r \leq \min \left(\frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha\xi(T)}{96D\psi_X^*\psi^2(1 + \frac{\alpha}{6\beta})^2} \right),$$

then

$$g_\gamma(\Delta_S, \Delta_L) \leq 2r.$$

Proof. The proof is very similar to the one given in Chandrasekaran, Parrilo & Willsky (2012) (suppl. mat. Prop. 3.2) but differs in a few different places. In order to spare the reader the trouble of going back and forth between Chandrasekaran, Parrilo & Willsky (2012) and this proof, we give all the details here, even for those parts that are identical.

Start by noticing that the objective function (B.2) is strictly convex: we showed earlier that the likelihood is convex; the tangent constraints ensure that it is *strictly* convex. As such, it has a unique minimum. Chandrasekaran, Parrilo & Willsky (2012) show that, by applying the optimality conditions of this function at $(\hat{S}_\Omega, \hat{L}_{T'})$, there exists two Lagrange multipliers $Q_{\Omega^\perp} \in \Omega^\perp$, $Q_{T'^\perp} \in T'^\perp$ such that

$$\begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\hat{S}_\Omega - \hat{L}_{T'}) + Q_{\Omega^\perp} \in \lambda_n \gamma \partial \|\hat{S}_\Omega\|_1,$$

and

$$\begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\hat{S}_\Omega - \hat{L}_{T'}) + Q_{T'^\perp} \in \lambda_n \partial \|\hat{L}_{T'}\|_*.$$

In particular, one checks that

$$\mathcal{P}_\Omega \left[\begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\hat{S}_\Omega - \hat{L}_{T'}) \right] = Z_\Omega,$$

and

$$\mathcal{P}_{T'} \left[\begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\hat{S}_\Omega - \hat{L}_{T'}) \right] = Z_{T'},$$

with $Z_\Omega \in \Omega$, $Z_{T'} \in T'$ and $\|Z_\Omega\|_\infty \leq \lambda_n \gamma$, $\|Z_{T'}\|_2 \leq \lambda_n$. Write $Z = (Z_\Omega, Z_{T'})$. We then have

$$\mathcal{P}_Y \mathcal{A}^\dagger \left[\begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\hat{S}_\Omega - \hat{L}_{T'}) \right] = Z,$$

with $g_\gamma(Z) \leq 2\lambda_n$. Moreover, because the optimum is unique, $(\hat{S}_\Omega, \hat{L}_{T'})$ is the unique solution to this equation.

We now rewrite the difference

$$\begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\hat{S}_\Omega - \hat{L}_{T'})$$

in terms of the errors (Δ_S, Δ_L) .

To that end remark that, by definition, $K^* = S^* - L^*$, so that

$$\begin{aligned} \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(\hat{S}_\Omega - \hat{L}_{T'}) &= \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(K^* + \mathcal{A}(\Delta_S, \Delta_L)) \\ &= \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}K^* - R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A}(\Delta_S, \Delta_L) \\ &\triangleq E_n - R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A}(\Delta_S, \Delta_L) \\ &= E_n - R_{K^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{AP}_Y(\Delta_S, \Delta_L) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{C}_{T'}. \end{aligned}$$

Now, since T' is such that $\rho(T, T') \leq \frac{\xi(T)}{2}$, we have from Proposition 1 that the operator $\mathcal{B} = (\mathcal{P}_Y \mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z^n}^* \mathcal{AP}_Y)^{-1}$ is an homeomorphism from $\mathcal{Y} \rightarrow \mathcal{Y}$.

Following the construction of Chandrasekaran, Recht, Parrilo & Willsky (2012),

we consider the following function

$$F: \mathcal{Y} \rightarrow \mathcal{Y}$$

$$(\delta_S, \delta_L) \mapsto (\delta_S, \delta_L) - \mathcal{B} \left(\mathcal{P}_Y \mathcal{A}^\dagger \left[E_n - R_{K^*}(\mathcal{A}(\delta_S, \delta_L)) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A} \mathcal{P}_Y(\delta_S, \delta_L) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{C}_{T'} \right] - Z \right).$$

Now, assume for a minute that there exist a fixed point (δ_S, δ_L) of F . By construction of F , (δ_S, δ_L) is a fixed point if and only if

$$0 = \mathcal{B} \left(\mathcal{P}_Y \mathcal{A}^\dagger \left[E_n - R_{K^*}(\mathcal{A}(\delta_S, \delta_L)) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A} \mathcal{P}_Y(\delta_S, \delta_L) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{C}_{T'} \right] - Z \right).$$

Since $\mathcal{B}$ is bijective, we conclude that (δ_S, δ_L) is a fixed-point if and only if

$$\mathcal{P}_Y \mathcal{A}^\dagger \left[E_n - R_{K^*}(\mathcal{A}(\delta_S, \delta_L)) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A} \mathcal{P}_Y(\delta_S, \delta_L) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{C}_{T'} \right] = Z.$$

Going back to the definition of Z given above, we see that, by construction, the fixed-point must be $\mathcal{P}_Y(\Delta_S, \Delta_L)$. Moreover, recall that the solution to that equation is unique.

We now focus on showing that such a fixed-point exists. In particular, we show that this unique fixed-point lies in

$$\mathbb{B}_r = \{(\delta_S, \delta_L) | g_\gamma(\delta_S, \delta_L) \leq r, (\delta_S, \delta_L) \in \mathcal{Y}\}.$$

To that end, let us show that $\mathcal{B}$ maps this closed ball onto itself and conclude the existence of a fixed-point by Brouwer's theorem.

First of all, F simplifies to

$$\mathcal{B} \left(\mathcal{P}_Y \mathcal{A}^\dagger \left[-E_n + R_{K^*}(\mathcal{A}(\delta_S, \delta_L)) - \mathcal{I}_{\Sigma_Z^n}^* \mathcal{C}_{T'} \right] + Z \right)$$

Now, we can apply Proposition 1 once more:

$$\begin{aligned}
g_\gamma(\delta_S, \delta_L) &\leq \frac{2}{\alpha} g_\gamma \left(\mathcal{P}_Y \mathcal{A}^\dagger \left[E_n - R_{K^*}(\mathcal{A}(\delta_S, \delta_L)) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{C}_{T'} \right] - Z \right) \\
&\leq \frac{4}{\alpha} \left(g_\gamma \left(\mathcal{A}^\dagger \left[E_n - R_{K^*}(\mathcal{A}(\delta_S, \delta_L)) + \mathcal{I}_{\Sigma_Z^n}^* \mathcal{C}_{T'} \right] \right) + \lambda_n \right) \\
&\leq \frac{r}{2} + \frac{4}{\alpha} g_\gamma \left(\mathcal{A}^\dagger R_{K^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) \right).
\end{aligned}$$

The second inequality uses the fact that $g_\gamma(\mathcal{P}_Y(\cdot, \cdot)) \leq 2g_\gamma(\cdot, \cdot)$, while the last one uses the definition of r .

Let us now bound the second summand of the last inequality, $\frac{4}{\alpha} g_\gamma \left(\mathcal{A}^\dagger R_{K^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) \right)$, by $\frac{r}{2}$. Remark that (δ_S, δ_L) since it is in $\mathbb{B}$. In addition, by our assumptions on r , we can apply Proposition 2:

$$\begin{aligned}
\frac{4}{\alpha} g_\gamma \left(\mathcal{A}^\dagger R_{K^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) \right) &\leq \frac{12D\psi_X^* \psi^2 (1 + \frac{\alpha}{6\beta})^2 g_\gamma(\delta_S, \delta_L)^2}{\alpha \xi(T)} \\
&\leq \frac{48D\psi_X^* \psi^2 (1 + \frac{\alpha}{6\beta})^2 r^2}{\alpha \xi(T)} \\
&\leq \frac{48D\psi_X^* \psi^2 (1 + \frac{\alpha}{6\beta})^2 r}{\alpha \xi(T)} \frac{\alpha \xi(T)}{96D\psi_X^* \psi^2 (1 + \frac{\alpha}{6\beta})^2} \\
&\leq \frac{r}{2}.
\end{aligned}$$

To go from the first line to the second, we used our assumptions on r and the fact that $g_\gamma(\delta_S, \delta_L)$ lies in the ball $\mathbb{B}_r$. Going from line 2 to 3 also uses our assumptions on r . Combining this result with the previous sequence of inequalities, we have shown that $F(\delta_S, \delta_L) \leq r$, for any (δ_S, δ_L) in $\mathbb{B}_r$. It follows that F maps the closed ball $\mathbb{B}_r$ onto itself and we are now within the range of applicability of Brouwer's fixed-point theorem. Subsequently, F admits a fixed-point within $\mathbb{B}_r$ and, as argued earlier, this fixed-point must be $\mathcal{P}_Y(\Delta_S, \Delta_L)$. Therefore, $g_\gamma(\mathcal{P}_Y(\Delta_S, \Delta_L)) \leq r$.

Finally, we have

$$\begin{aligned} g_\gamma(\Delta_S, \Delta_L) &\leq g_\gamma(\mathcal{P}_Y(\Delta_S, \Delta_L)) + \|\mathcal{C}_{T'}\| + 2 \\ &\leq 2r. \end{aligned}$$

□

B.5.3 Variety constraints

We now turn to a variant of the problem in which both the rank and the sparsity pattern are enforced explicitly. This amounts to minimising the objective function within the following non-convex constraint set:

$$\begin{aligned} \mathcal{M} = \{ (S, L) | S \in \Omega(S^*), \text{rank}(L) \leq \text{rank}(L^*), \\ \|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right)}, \\ g_\gamma(\mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z^n}^* \mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n \}. \end{aligned}$$

We denote by $(\hat{S}_M, \hat{L}_M)$ the solution to our problem under these constraints, *i.e.*:

$$\begin{aligned} (\hat{S}_M, \hat{L}_M) &= \arg \min_{S, L} -\log \det(S_X - L_X) + \text{tr}(\Sigma_X^n(S_X - L_X)) + 2\text{tr}(\Sigma_{ZX}^n(S_{ZX} + L_{ZX})^T) \\ &\quad + \text{tr}(\Sigma_Z^n(S_{ZX} + L_{ZX})(S_X - L_X)^{-1}(S_{ZX} + L_{ZX})^T) \\ &\quad \text{s.t. } S_X - L_X \succ 0, (S, L) \in \mathcal{M} \\ &\quad \text{where } S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}; \end{aligned} \tag{B.3}$$

We start by proving a preliminary result that bounds the spectral norm of the FIM evaluated at the nominal parameters. This simple bound will be used later in this section.

Proposition 7. Recall that $K^* = \begin{pmatrix} K_X^* \\ K_{ZX}^* \end{pmatrix}$ and $\mathcal{I}_{\Sigma_Z^n}^* \triangleq \mathcal{I}_{\Sigma_Z^n}(K^*)$.

$$\|\mathcal{I}_{\Sigma_Z^n}^*\|_2 \leq \psi_X^{*2} \left(1 + 2 \frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2 \right).$$

Proof. As shown earlier, we have

$$\mathcal{I}_{\Sigma_Z^n}^* = - \begin{pmatrix} K_X^{*-1} \otimes K_X^{*-1} & 0 \\ 0 & 0 \end{pmatrix} - 2K_X^{*-1} \otimes \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^n K_{ZX}^* K_X^{*-1} & -K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^n \\ \cdot & \Sigma_Z^n \end{pmatrix}.$$

Therefore,

$$\begin{aligned} \|\mathcal{I}_{\Sigma_Z^n}^*\|_2 &\leq \psi_X^{*2} + 2\psi_X^* \left\| \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^n K_{ZX}^* K_X^{*-1} & -K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^n \\ \cdot & \Sigma_Z^n \end{pmatrix} \right\|_2 \\ &= \psi_X^{*2} + 2\psi_X^* \left\| \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \sqrt{\Sigma_Z^n} \\ \sqrt{\Sigma_Z^n} \end{pmatrix} \begin{pmatrix} K_X^{*-1} K_{ZX}^{*T} \sqrt{\Sigma_Z^n} \\ \sqrt{\Sigma_Z^n} \end{pmatrix}^T \right\|_2 \\ &\leq \psi_X^{*2} + 2\psi_X^* (\psi_X^* \phi_{ZX}^* \sqrt{\psi_Z} + \sqrt{\psi_Z})^2 \\ &\leq \psi_X^{*2} (1 + 2 \frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2). \end{aligned}$$

□

We now have the following results. Remark that because we derived different bounds on the curvature of the rank variety and on the spectral norm of $\mathcal{I}_{\Sigma_Z^n}^*$, the assumptions of these propositions differ from the one made in Chandrasekaran, Parrilo & Willsky (2012). However, given the above bound on $\|\mathcal{I}_{\Sigma_Z^n}^*\|_2$, the proof is a straightforward adaptation of the proof in Chandrasekaran, Parrilo & Willsky (2012).

Proposition 8. (see Chandrasekaran, Parrilo & Willsky (2012) (suppl. mat.), Proposition 3.3)

Consider any $(S, L) \in \mathcal{M}$ and let $(\Delta_S, \Delta_L) = (S - S^*, L^* - L)$. For γ in the range

required for consistency and letting $C_1 = \frac{48}{\alpha} + \frac{1}{\psi_X^{*2} \left(1 + 2 \frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2 \right)}$, we have that $g_\gamma(\Delta_S, \Delta_L) \leq C_1 \lambda_n$.

The following corollary shows that, under some suitable conditions, any solution to (B.3) is algebraically consistent. As a matter of fact, it holds for any pair in $\mathcal{M}$, irrespective of whether it is a solution to (B.3).

Here again, the proof of that corollary follows straightforwardly from the one given in Chandrasekaran, Parrilo & Willsky (2012). However, unlike the previous proposition, the statement of the following corollary does not obviously follow from Chandrasekaran, Parrilo & Willsky (2012) (suppl. mat.), Corollary 3.4. For that reason, we give details.

Corollary 3. *(see Chandrasekaran, Parrilo & Willsky (2012) (suppl. mat.), Corollary 3.4)*

Consider any pair $(S, L) \in \mathcal{M}$ and, as before, let $\Delta_S = S - S^*$, $\Delta_L = L^* - L$. For C_1 defined as in the previous proposition, further define the following constants:

- $C_2 = \left(1 + \frac{24(2-\nu)}{\nu} \right) C_1^2 D \psi_X^{*2} \left(1 + 2 \frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2 \right)$;
- $C_3 = C_1 + \frac{3\alpha C_1^2(2-\nu)}{4(3-\nu)}$;
- $\sigma = \text{smallest non-zero singular value of } L^*$;
- $C_4 = \max\{C_2, C_3\}$;
- $C_5 = \frac{C_1 \nu \alpha}{\beta(2-\nu)}$;
- $T' = T(L)$ and $\mathcal{C}_{T'} = \mathcal{P}_{T'^\perp}(L^*)$.

Assume that $\sigma \geq \frac{C_4 \lambda_n}{\xi(T)^2}$, and suppose that the smallest entry in magnitude of S^* is greater than $\frac{C_5 \lambda_n}{\mu(\Omega)}$. Further assume that γ is in the acceptable range defined earlier. Then we have:

1. L has rank equal to L^* , and $L \succeq 0$;

2. $\text{sign}(S) = \text{sign}(S^*)$;

3. $\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \leq \frac{4\xi(T)\lambda_n}{73D\psi_X^{*2}\left(1+2\frac{\psi_Z}{\psi_X^*}(1+\psi_X^*\phi_{ZX}^*)^2\right)}$.

4. $\rho(T, T') \leq \frac{8\xi(T)}{73}$.

5. $g_\gamma(\mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z}^* \mathcal{C}_{T'}) \leq \frac{\lambda_n \nu}{6(2-\nu)}$.

6. $\|\mathcal{C}_{T'}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$.

Proof. First of all, remark that

$$\begin{aligned} C_1 &\geq \frac{1}{\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right)} \\ &\geq \frac{1}{w} \frac{1}{\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right)} \\ &\geq \frac{\xi(T)}{D} \frac{1}{\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right)}. \end{aligned}$$

We also have $\xi(T) \leq 1$ and, by Proposition 3, $\|\Delta_L\|_2 \leq C_1 \lambda_n$. One can also check that, for $0 < \nu \leq \frac{1}{2}$, $\frac{24(2-\nu)}{\nu} \geq 72$.

By our assumptions,

$$\begin{aligned} \sigma &\geq \frac{C_4 \lambda_n}{\xi(T)^2} \\ &\geq \frac{C_2 \lambda_n}{\xi(T)^2} \\ &\geq \frac{73C_1^2 D \psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right) \lambda_n}{\xi(T)^2} \\ &\geq \frac{73C_1 \lambda_n}{\xi(T)} \\ &\geq 8C_1 \lambda_n \\ &\geq 8\|\Delta_L\|_2. \end{aligned}$$

So, the smallest singular value of the nominal low-rank piece, L^* is greater than $4\|L - L^*\|_2$. Consequently, L and L^* have the same rank. Moreover, since $\|L^*\|_2 \geq \max(\|L_X^*\|_2, \|L_{ZX}^*\|_2)$, $\sigma \leq$ (smallest singular value of L_X^*). Let us call σ_X the smallest eigenvalue of L_X^* . On the other hand, $\|\Delta_L\| \geq \max(\|\Delta_{L_X}\|_2, \|\Delta_{L_{ZX}}\|_2)$, so that

$$\sigma_X \geq \sigma \geq 8\|\Delta_L\|_2 \geq 8\|\Delta_{L_X}\|_2.$$

This implies that L_X and L_X^* have the same inertia, *i.e.* the same number of positive, null and negative eigenvalues.

This proves 1.

Now, since $\sigma > 4\|\Delta_L\|_2$, we can apply Proposition 2 and therefore

$$\begin{aligned} \|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 &\leq \frac{4\|\Delta_L\|_2^2}{\sigma} \\ &\leq \frac{4C_1^2\xi(T)^2\lambda_n^2}{C_2\lambda_n} \\ &\leq \frac{4\xi(T)^2\lambda_n}{73D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^*\phi_{ZX}^*)^2\right)} \\ &\leq \frac{4\xi(T)\lambda_n}{73D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^*\phi_{ZX}^*)^2\right)}. \end{aligned}$$

This proves 3.

Now, we apply the other part of Proposition 2, so that

$$\begin{aligned} \rho(T, T') &\leq \frac{8\|\Delta_L\|_2}{\sigma} \\ &\leq \frac{8C_1\xi(T)^2}{C_2\lambda_n} \\ &\leq \frac{8\xi(T)^2}{73C_1D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^*\phi_{ZX}^*)^2\right)} \\ &\leq \frac{8\xi(T)}{73}. \end{aligned}$$

This proves 4.

Let σ' be the smallest non-zero singular value of L .

$$\begin{aligned}
\sigma' &\geq \frac{C_2\lambda_n}{\xi(T)^2} - C_1\lambda_n \\
&\geq C_1\lambda_n \left(\frac{73C_1D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^*\phi_{ZX}^*)^2\right)}{\xi(T)^2} - 1 \right) \\
&\geq 5\|\Delta_L\|_2.
\end{aligned}$$

Here again, let us apply Proposition 2 but this time with σ' :

$$\begin{aligned}
\|\mathcal{C}_{T'}\|_2 &\leq \frac{4\|\Delta_L\|_2^2}{\sigma'} \\
&\leq \frac{4C_1^2\lambda_n^2}{\frac{C_2\lambda_n}{\xi(T)^2} - C_1\lambda_n} \\
&= \frac{4C_1^2\xi(T)^2\lambda_n}{\frac{24(2-\nu)}{\nu}C_1^2D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^*\phi_{ZX}^*)^2\right) + C_1^2D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^*\phi_{ZX}^*)^2\right) - C_1\xi(T)^2} \\
&\leq \frac{4C_1^2\xi(T)^2\lambda_n}{\frac{24(2-\nu)}{\nu}C_1^2D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^*\phi_{ZX}^*)^2\right)} \\
&\leq \frac{\nu\xi(T)\lambda_n}{6(2-\nu)D\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^*\phi_{ZX}^*)^2\right)},
\end{aligned}$$

so that

$$g_\gamma(\mathcal{A}^\dagger \mathcal{I}_{\Sigma_Z^n}^* \mathcal{C}_{T'}) \leq w\|\mathcal{I}_{\Sigma_Z^n}^*\|_2\|\mathcal{C}_{T'}\|_2 \leq \frac{\nu\lambda_n}{6(2-\nu)}.$$

This proves 5.

Proof for 6. and 2. still remain. We have that

$$\begin{aligned}
\sigma' &\geq \frac{C_3\lambda_n}{\xi(T)^2} - C_1\lambda_n \\
&\geq \frac{3\alpha C_1^2(2-\nu)}{4(3-\nu)}\lambda_n,
\end{aligned}$$

which can be used to prove the desired bound on $\|\mathcal{C}_{T'}\|_2$. As before, we have

$$\begin{aligned}\|\mathcal{C}_{T'}\|_2 &\leq \frac{4C_1^2\lambda_n^2}{\sigma'} \\ &\leq \frac{4C_1^2\lambda_n^2}{\frac{3\alpha C_1^2(2-\nu)}{4(3-\nu)}\lambda_n} \\ &\leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}.\end{aligned}$$

This proves 6.

Finally, in order to prove 2., recall that $\|\Delta_S\|_\infty \leq \gamma C_1 \lambda_n$. In order to show that $\text{sign}(S^*) = \text{sign}(S)$ we need to show that the magnitude of the smallest entry of S^* is greater $\|\Delta_S\|_\infty$. Since γ is in the acceptable range, we know that

$$\gamma \leq \frac{\nu\alpha}{2\beta\mu(\Omega)(2-\nu)} \leq \frac{\nu\alpha}{\beta\mu(\Omega)(2-\nu)}.$$

So that, by the assumption on the magnitude of the smallest entry of S^* , θ , we have

$$\theta \geq \frac{C_1\nu\alpha\lambda_n}{\mu(\Omega)\beta(2-\nu)} \geq C_1\lambda_n\gamma.$$

□

B.5.4 From variety to tangent space constraints

In the previous section, we showed that any solution to the non-convex problem (B.3) was algebraically consistent (among other properties). This problem was non-convex due to the constraint $\text{rank}(L) \leq \text{rank}(L^*)$. Here, we consider a different function where the constraint on the rank of L is replaced by another one $L \in T_{\mathcal{M}} \triangleq T(\hat{L}_{\mathcal{M}})$, where $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$ is any solution to (B.3).

We denote by $(\hat{S}_{\mathcal{M}}, \hat{L}_{T_{\mathcal{M}}})$ the solution to that modified convex problem.

The following proposition shows that, under suitable conditions, the solution to

that modified problem is the same as the variety constrained one. In particular, it is algebraically consistent.

The proof is lengthy and technical. However, we tailored the assumptions of the previous results (the definitions of C_2, C_3 in the previous corollary in particular) so as to achieve bounds that are identical – or tighter – than the ones derived in Chandrasekaran, Parrilo & Willsky (2012) (Prop. 3.5 of the suppl. materials). We therefore refer the reader to Chandrasekaran, Parrilo & Willsky (2012) for the proof. Given the result we proved in the previous section, understanding their proof does not raise any challenges.

Proposition 9. *(see Chandrasekaran, Parrilo & Willsky (2012) (suppl. mat.), Proposition 3.5)*

Let γ be in the correct range. Suppose that the minimum non-zero singular value of L^* is such $\sigma \geq \frac{C_4 \lambda_n}{\xi(T)^2}$, and suppose that the smallest entry in magnitude of S^* is greater than $\frac{C_5 \lambda_n}{\mu(\Omega)}$ (with $C_{4,5}$ defined as above). Let $g_\gamma(\mathcal{A}^\dagger E^n) \leq \frac{\lambda_n \nu}{6(2-\nu)}$. Further suppose that

$$\lambda_n \leq \frac{3\alpha(2-\nu)}{16(3-\nu)} \min \left(\frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha\xi(T)}{96D\psi_X^*\psi^2(1+\frac{\alpha}{6\beta})^2} \right).$$

Then we have $(\hat{S}_\Omega, \hat{L}_{T_M}) = (\hat{S}_M, \hat{L}_M)$.

Corollary 4. *Under the assumptions of Proposition 9, we have that $\text{rank}(\hat{L}_{T_M}) = \text{rank}(L^*)$ and that $T(\hat{L}_{T_M}) = T_M$. Similarly, $\text{sign}(\hat{S}_\Omega) = \text{sign}(S^*)$.*

B.5.5 From tangent space constraints to problem (B.1)

At this stage, we know under which conditions the problem with tangent space constraints is algebraically consistent. The following result derives a condition under which this constrained problem admits the same solution as (B.1). Recall that (B.1) differs from the original problem by the removal of the constraint $L_X \succeq 0$.

Lemma 1. (see, Chandrasekaran, Parrilo & Willsky (2012) (suppl. mat.), Lemma 3.7)

Let $(\hat{S}_\Omega, \hat{L}_{T_M})$ be defined as in the previous section. Suppose that the assumptions of Proposition 9 hold. Further assume that

$$g_\gamma(\mathcal{A}^\dagger R_{K^*} \mathcal{A}(\Delta_S, \Delta_L)) \leq \frac{\lambda_n \nu}{6(2 - \nu)}.$$

Then $(\hat{S}_\Omega, \hat{L}_{T_M})$ is also the unique optimum to problem (B.1).

B.5.6 Bounding the error terms

We start by recalling a useful result from the literature. This result gives insight into the non-asymptotic behaviour of random matrices.

Theorem 6. (Davidson & Szarek (2001), Theorem II.13)

Given $n, p \in \mathbb{N}$, with $p \leq n$, let Γ be $p \times n$ matrix with i.i.d. drawn from a normal distribution with mean 0 and variance $1/n$. Then the largest and smallest singular values of Γ , $\sigma_1(\Gamma)$ and $\sigma_p(\Gamma)$, are such that

$$\max \left(\mathbb{P} \left(\sigma_1(\Gamma) \geq 1 + \sqrt{\frac{p}{n}} + t \right), \mathbb{P} \left(\sigma_p(\Gamma) \leq 1 - \sqrt{\frac{p}{n}} - t \right) \right) \leq \exp \left(-\frac{nt^2}{2} \right),$$

for $t > 0$.

Our goal is to bound the spectral norm of the error terms:

$$E^n = \begin{pmatrix} \Sigma_X^n \\ 2\Sigma_{ZX}^n \end{pmatrix} - \mathcal{F}_{\Sigma_Z^n}(K^*).$$

We recall the notations defined in Proposition 5: we write $\psi_Z \triangleq \|\Sigma_Z^n\|_2$, $\psi_X^* \triangleq \|K_X^{*-1}\|_2$ and $\phi_{ZX}^* \triangleq \|K_{ZX}^*\|_2$.

Before proving a result about $\|E^n\|_2$, we start by deriving bounds on $\|E_{ZX}^n\|_2$ and $\|E_X^n\|_2$ independently.

Proposition 10. *Let $\|E_{ZX}^n\|_2$ be defined as above. Then, given any $\delta > 0$ and for $n \geq \frac{16\psi_Z\psi_X^*}{\delta^2}(\sqrt{m} + \sqrt{p})^2$,*

$$\mathbb{P}(\|E_{ZX}^n\|_2 \geq \delta) \leq \exp\left(-\frac{n\delta^2}{32\psi_Z\psi_X^*}\right).$$

Proof. Recall that E_{ZX}^n was defined as:

$$E_{ZX}^n = 2(\Sigma_{ZX}^n + 2\Sigma_Z K_{ZX}^* K_X^{*-1}).$$

Recall also that the samples Y_X are drawn conditionally on Y_Z according to:

$$Y_X|Y_Z \sim \mathcal{N}(-K_X^{*-1}K_{ZX}^{*T}Y_Z, K_X^{*-1}),$$

so that

$$Y_X = -Y_Z K_{ZX}^* K_X^{*-1} + R$$

, with R a matrix whose rows are drawn according to $R_i \sim \mathcal{N}(0, K_X^{*-1}), \forall 1 \leq i \leq n$.

It follows that

$$\frac{n}{2}E_{ZX}^n = Y_Z^T Y_X + Y_Z^T Y_Z K_{ZX}^* K_X^{*-1} = Y_Z^T R.$$

Thus, bounding $\|E_{ZX}^n\|_2$ amounts to bounding the largest singular value of $\frac{2}{n}Y_Z^T R$.

For convenience, we write $p_+ = \max(m, p)$ and $p_- = \min(m, p)$.

Now, because the rows of R all have covariance matrix K_X^{*-1} we have

$$\begin{aligned} \mathbb{P}(\|E_{ZX}^n\|_2 \geq \delta) &= \mathbb{P}\left(\left\|\frac{2}{n}Y_Z^T R\right\|_2 \geq \delta\right) \\ &\leq \mathbb{P}\left(\left\|\frac{1}{\sqrt{n}}A\right\|_2 \geq \frac{\delta}{2\sqrt{\psi_Z\psi_X^*}}\right) \end{aligned}$$

where A is a $p_- \times p_+$ matrix whose entries are i.i.d. drawn from a standard normal.

Thus, $\frac{\sqrt{1}}{\sqrt{p_+}}A$ satisfies Theorem 6. Therefore, we have

$$\begin{aligned}\mathbb{P}(\|E_{ZX}^n\|_2 \geq \delta) &\leq \mathbb{P}\left(\left\|\frac{1}{\sqrt{n}}A\right\|_2 \geq \frac{\delta}{2\sqrt{\psi_Z\psi_X^*}}\right) \\ &\leq \mathbb{P}\left(\left\|\frac{1}{\sqrt{p_+}}A\right\|_2 \geq \frac{\delta\sqrt{n}}{2\sqrt{\psi_Z\psi_X^*p_+}}\right).\end{aligned}$$

Applying Theorem 6 to $\frac{1}{\sqrt{p_+}}A$, we obtain

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{p_+}}A\right\|_2 \geq 1 + \sqrt{\frac{p_-}{p_+}} + t\right) \leq \exp\left(-\frac{p_+t^2}{2}\right),$$

so that

$$\mathbb{P}\left(\|E_{ZX}^n\|_2 \geq \frac{2\sqrt{\psi_Z\psi_X^*p_+}}{\sqrt{n}} + \frac{2\sqrt{\psi_Z\psi_X^*p_-}}{\sqrt{n}} + t\right) \leq \exp\left(-\frac{nt^2}{8\psi_Z\psi_X^*}\right).$$

By our assumptions on n , we know that $\frac{\delta}{2} \geq \frac{2\sqrt{\psi_Z\psi_X^*}}{\sqrt{n}}(\sqrt{p_-} + \sqrt{p_+})$. We conclude the proof by picking $t = \delta/2$. $\square$

We now turn to $\|E_X^n\|_2$ and prove a similar result.

Proposition 11. (see Chandrasekaran, Parrilo & Willsky (2012), suppl. materials Lemma 3.9)

Let E_X^n be defined as above. Given any $\delta > 0$, with $\delta \leq 8\psi_X^*$ and $n \geq \frac{64p\psi_X^{*2}}{\delta^2}$. Then,

$$\mathbb{P}(\|E_X^n\|_2 \geq \delta) \leq 2\exp\left(-\frac{n\delta^2}{128\psi_X^{*2}}\right).$$

Proof. We have:

$$E_X^n = \Sigma_X^n - (K_X^{*-1} + K_X^{*-1} K_{ZX}^{*T} \Sigma_Z^n K_{ZX}^* K_X^{*-1}).$$

Just like in the previous proposition, we can rewrite the error term in terms of R , an $n \times p$ matrix whose rows are drawn from $\mathcal{N}(0, K_X^{*-1})$:

$$E_X^n = \frac{1}{n} R^T R - K_X^{*-1}.$$

The rest of the proof follows a progression which is identical to Chandrasekaran, Parrilo & Willsky (2012) (suppl. materials, Lemma 3.9). $\square$

We now put these propositions together in order to bound $\|E^n\|_2$.

Proposition 12. *Let E^n be defined as above. For any $\delta > 0$ with $\delta \leq 8\sqrt{2}\psi_X^*$, let $n \geq \frac{1}{\delta^2} \max(128p\psi_X^{*2}, 32\psi_Z\psi_X^*(\sqrt{m} + \sqrt{p})^2)$. Then,*

$$\mathbb{P}(\|E^n\|_2 \geq \delta) \leq \max\left(2 \exp\left(-\frac{n\delta^2}{256\psi_X^{*2}}\right), \exp\left(-\frac{n\delta^2}{64\psi_Z\psi_X^*}\right)\right).$$

Proof. We know that

$$\max(\|E_X^n\|_2, \|E_{ZX}^n\|_2) \leq \|E^n\|_2 \leq \sqrt{2} \max(\|E_X^n\|_2, \|E_{ZX}^n\|_2)$$

so that

$$\mathbb{P}(\|E^n\|_2 \geq \delta) \leq \max\left(\mathbb{P}(\sqrt{2}\|E_X^n\|_2 \geq \delta), \mathbb{P}(\sqrt{2}\|E_{ZX}^n\|_2 \geq \delta)\right).$$

From Proposition 10, we have

$$\mathbb{P}(\sqrt{2}\|E_{ZX}^n\|_2 \geq \delta) \leq \exp\left(-\frac{n\delta^2}{64\psi_Z\psi_X^*}\right),$$

for $n \geq \frac{32\psi_Z\psi_X^*}{\delta^2}(\sqrt{m} + \sqrt{p})^2$.

Applying Proposition 11, we have for $\delta \leq 8\sqrt{2}\psi_X^*$ and $n \geq \frac{128p\psi_X^{*2}}{\delta^2}$:

$$\mathbb{P}(\sqrt{2}\|E_X^n\|_2 \geq \delta) \leq 2 \exp\left(-\frac{n\delta^2}{256\psi_X^{*2}}\right).$$

□

Finally, we can prove a result that allows us to condition on the norm of the error terms as a function of the number of samples and the dimensions of the problem, p and m .

Corollary 5. *Let $M = \max\left(1, \frac{\psi_Z}{4\psi_X^*}(1 + \sqrt{\frac{m}{p}})^2\right)$. Let $\delta_n = \sqrt{\frac{256\psi_X^{*2}pM}{n}}$. Then for $n \geq 2pM$,*

$$\mathbb{P}(\|E^n\|_2 \leq \delta_n) \geq 1 - 2 \min\left(\exp(-pM), \exp\left(-\frac{4\psi_X^*}{\psi_Z}pM\right)\right).$$

Proof. We show that

$$\mathbb{P}(\|E^n\|_2 \geq \delta_n) \leq 2 \max(\exp(-pM), \exp(-\frac{4\psi_X^*}{\psi_Z}pM)).$$

First of all, remark that for $n \geq 2pM$, we have $\delta_n \leq 8\sqrt{2}\psi_X^*$, so we can apply Proposition 12.

Then, by Proposition 12, we have a bound of the form

$$\mathbb{P}(\|E^n\|_2 \geq \delta_n) \leq 2 \max(\exp(-A), \exp(-B)),$$

with

$$A = \frac{n}{256\psi_X^{*2}} \frac{256\psi_X^{*2}pM}{n} = pM,$$

and

$$B = \frac{n}{64\psi_Z\psi_X^*} \frac{256\psi_X^{*2}pM}{n} = \frac{4\psi_X^*}{\psi_Z} pM = \frac{4\psi_X^*}{\psi_Z} pM.$$

□

B.5.7 Proof of consistency

Before concluding with our final result, the consistency of estimator (2.1), we state a proposition which summarises the results of the previous sections, namely sections A.5.2, A.5.3, A.5.4 and A.5.5.

Proposition 13. *(Summary of the previous results)*

Define the following constants,

$$\begin{aligned} C_1 &= \frac{48}{\alpha} + \frac{1}{\psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right)}; \\ C_2 &= \left(1 + \frac{24(2-\nu)}{\nu}\right) C_1^2 D \psi_X^{*2} \left(1 + 2\frac{\psi_Z}{\psi_X^*} (1 + \psi_X^* \phi_{ZX}^*)^2\right); \\ C_3 &= C_1 + \frac{3\alpha C_1^2(2-\nu)}{4(3-\nu)}; \quad C_4 = \max\{C_2, C_3\}; \quad C_5 = \frac{C_1 \nu \alpha}{\beta(2-\nu)}; \end{aligned}$$

$\sigma = \text{smallest non-zero singular value of } L^*$;

$\theta = \text{smallest entry in magnitude of } S^*$.

Assume that γ is in the acceptable range, i.e.

$$\gamma \in \left[\frac{3\xi(T)\beta(2-\nu)}{\nu\alpha}, \frac{\nu\alpha}{2\mu(\Omega)\beta(2-\nu)} \right].$$

Assume that $\sigma \geq \frac{C_4\lambda_n}{\xi(T)^2}$ and $\theta \geq \frac{C_5\lambda_n}{\mu(\Omega)}$. Further assume that $g_\gamma(\mathcal{A}^\dagger E^n) \leq \frac{\lambda_n \nu}{6(2-\nu)}$ and

$$\lambda_n \leq \frac{3\alpha(2-\nu)}{16(3-\nu)} \min \left(\frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha\xi(T)}{96D\psi_X^*\psi^2(1 + \frac{\alpha}{6\beta})^2} \right).$$

Then, there exists a T' and a corresponding unique solution $(\hat{S}_\Omega, \hat{L}_{T'})$ to (B.2) with the following properties:

1. $\text{sign}(\hat{S}_\Omega) = \text{sign}(S^*)$, $\text{rank}(\hat{L}_{T'}) = \text{rank}(L^*)$ and $(\hat{L}_{T'})_X \succeq 0$. Moreover, $T(\hat{L}_{T'}) = T'$ and $\rho(T, T') \leq \frac{8\xi(T)}{73} \leq \frac{\xi(T)}{4}$.
2. As before, write $\mathcal{C}_{T'} = \mathcal{P}_{T'^\perp}(L^*)$. Then we have $g_\gamma(\mathcal{A}^\dagger \mathcal{I}_{\Sigma_{\mathbb{Z}}}^* \mathcal{C}_{T'}) \leq \frac{\lambda_n \nu}{6(2-\nu)}$ and $\|\mathcal{C}_{T'}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$.

In addition, if

$$g_\gamma(\mathcal{A}^\dagger R_{K^*} \mathcal{A}(\Delta_S, \Delta_L)) \leq \frac{\lambda_n \nu}{6(2-\nu)},$$

then $(\hat{S}_\Omega, \hat{L}_{T'}) = (\hat{S}, \hat{L})$, the unique solution to the original problem (2.1) (including the constraint $L_X \succeq 0$).

Our goal is now to combine the results of the previous section (bound on the error terms) in order to show under which conditions the assumptions of Proposition 13 are satisfied. We use Proposition 13 and Corollary 5.

We reuse the notations and constants introduced throughout the proof and refer the reader to the statement of Theorem 5 for a list of their definitions.

We define the list of assumptions required to achieve algebraic consistency.

Assumption 3. (*Assumptions for Algebraic Consistency*)

1. Set

$$C_6 = \frac{\alpha \nu}{32(3-\nu)D} \min \left(\frac{1}{6\psi_X^*}, \frac{\phi_{ZX}^*}{4}, \frac{\alpha \nu}{384D(3-\nu)\psi_X^* \psi^2(1 + \frac{\alpha}{6\beta})^2} \right)$$

and, for M defined as in Corollary 5., let n be such that

$$n \geq \frac{pM}{\xi(T)^4} \max(2, \frac{256\psi_X^{*2}}{C_6^2}).$$

2. For M defined as before, set

$$\delta_n = \sqrt{\frac{256\psi_X^{*2}pM}{n}},$$

and let

$$\lambda_n = \frac{6D\delta_n(2-\nu)}{\xi(T)\nu}.$$

3. Let the minimum non-zero singular value of L^* be such that

$$\sigma \geq \frac{C_4\lambda_n}{\xi(T)^2}.$$

4. Let the minimum magnitude nonzero entry θ of S^* be such that

$$\theta \geq \frac{C_5\lambda_n}{\mu(\Omega)}.$$

Assuming that Assumption 3 holds, the proof of our main Theorem 2 (which is restated for convenience in Theorem 5) is in all points identical to the proof given in Chandrasekaran, Parrilo & Willsky (2012), Section 5.5. We refer the reader to their proof for the details. In particular, the proof makes use of Proposition 6 in order to show that $g_\gamma(\Delta_S, \Delta_L) \leq \frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)}$.

Bibliography

Aguet, F. et al. (2016), Local genetic effects on gene expression across 44 human tissues, Technical report.

URL: <http://dx.doi.org/10.1101/074450>

Angrist, J. D. & Imbens, G. W. (1995), ‘Two-stage least squares estimation of average causal effects in models with variable treatment intensity’, *Journal of the American Statistical Association* **90**(430), 431–442.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996), ‘Identification of causal effects using instrumental variables’, *Journal of the American Statistical Association* **91**(434), 444–455.

Ashburner, M. et al. (2000), ‘Gene ontology: tool for the unification of biology’, *Nature Genetics* **25**(1), 25–29.

Bach, F. (2008), ‘Consistency of trace norm minimization’, *Journal of Machine Learning Research* **8**, 1019–1048.

Battle, A. et al. (2013), ‘Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals’, *Genome Research* **24**(1), 14–24.

Boyd, S. (2011), ‘Alternating direction method of multipliers’, *Proceedings of the 51st IEEE Conference on Decision and Control*.

- Bühlmann, P. & Van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, 1st edn, Springer Publishing Company, Incorporated.
- Candès, E. J., Li, X., Ma, Y. & Wright, J. (2011), ‘Robust principal component analysis?’, *J. ACM* **58**(3), 1–37.
- Chandrasekaran, V., Parrilo, P. A. & Willsky, A. S. (2012), ‘Latent variable graphical model selection via convex optimization’, *The Annals of Statistics* **40**(4), 1935–1967.
- Chandrasekaran, V., Recht, B., Parrilo, P. A. & Willsky, A. S. (2012), ‘The convex geometry of linear inverse problems’, *Foundations of Computational Mathematics* **12**(6), 805–849.
- Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. & Willsky, A. S. (2009), ‘Rank-sparsity incoherence for matrix decomposition’, *SIAM Journal on Optimization* **21**(2), 572–596.
- Chen, J. & Chen, Z. (2008), ‘Extended bayesian information criteria for model selection with large model spaces’, *Biometrika* .
- Chen, L., Davey Smith, G., Harbord, R. M. & Lewis, S. J. (2008), ‘Alcohol intake and blood pressure: a systematic review implementing a mendelian randomization approach.’, *PLoS medicine* **5**(3), e52.
- Chen, L. & Huang, J. (2016), ‘Sparse reduced-rank regression with covariance estimation’, *Statistics and Computing* **26**(1), 461–470.
- Chickering, D. M. & Boutilier, C. (2002), ‘Optimal structure identification with greedy search’, *Journal of Machine Learning Research* **3**, 507–554.

- Combettes, P. L. & Pesquet, J.-C. (2009), Proximal splitting methods in signal processing, *in* ‘Fixed-Point Algorithms for Inverse Problems in Science and Engineering’, chapter 10, pp. 185–212.
- Davidson, K. R. & Szarek, S. J. (2001), Handbook of the geometry of banach spaces, *in* W. B. Johnson & J. Lindenstrauss, eds, ‘Handbook of the Geometry of Banach Spaces’.
- Dawid, P. (2002), ‘Influence Diagrams for Causal Modelling and Inference’, *International Statistical Review / Revue Internationale de Statistique* **70**(2), 161–189.
- Fan, J. & Liu, H. (2013), ‘Statistical analysis of big data on pharmacogenomics’, *Advanced Drug Delivery Reviews* **65**(7), 987–1000.
- Foygel, R. & Drton, M. (2010), Extended bayesian information criteria for gaussian graphical models, *in* J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel & A. Culotta, eds, ‘Advances in Neural Information Processing Systems 23’, Curran Associates, Inc., pp. 604–612.
- Friedman, J., Hastie, T. & Tibshirani, R. (2008), ‘Sparse inverse covariance estimation with the graphical lasso’, *Biostatistics* **9**(3), 432–41.
- Gao, X., Pu, D. Q., Wu, Y. & Xu, H. (2012), ‘Tuning parameter selection for penalized likelihood estimation of gaussian graphical model’, *Statistica Sinica* **22**(3), 1123–1146.
- Hastings et al. (2012), ‘The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013’, *Nucleic Acids Research* **41**(D1), D456–D463.
- Heckman, J. J. & Robb, R. (1985), ‘Alternative methods for evaluating the impact of interventions’, *Journal of Econometrics* **30**(1), 239–267.

- Hernán, M. A. & Robins, J. M. (2006), ‘Instruments for causal inference: an epidemiologist’s dream?’, *Epidemiology (Cambridge, Mass.)* **17**(4), 360–372.
- Johnstone, I. M. (2001), ‘On the distribution of the largest eigenvalue in principal components analysis’, *Ann. Statist.* **29**(2), 295–327.
- Kang, H., Zhang, A., Cai, T. T. & Small, D. S. (2016), ‘Instrumental variables estimation with some invalid instruments and its application to mendelian randomization’, *Journal of the American Statistical Association* **111**(513), 132–144.
- Koller, D. & Friedman, N. (2009), *Probabilistic Graphical Models: Principles and Techniques*, The MIT Press.
- Lauritzen, S. (1996), *Graphical Models*, Clarendon Press.
- Löfberg, J. (2004), Yalmip : A toolbox for modeling and optimization in matlab, in ‘Proceedings of the CACSD Conference’, pp. 284 – 289.
- Maathuis, M. H., Kalisch, M. & Bühlmann, P. (2009), ‘Estimating high-dimensional intervention effects from observational data’, *The Annals of Statistics* **37**(6A), 3133–3164.
- Meinshausen, N. & Bühlmann, P. (2010), ‘Stability selection’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **72**(4), 417473.
- Mostafavi, S., Battle, A., Zhu, X., Urban, A. E., Levinson, D., Montgomery, S. B. & Koller, D. (2013), ‘Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge’, *PLoS ONE* **8**(7), e68141.
- Nesterov, Y. (2005), ‘Smooth minimization of non-smooth functions’, *Mathematical Programming* **103**(1), 127–152.

- Ng, A. Y. & Jordan, M. I. (2002), On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, *in* ‘Advances in Neural Information Processing Systems 14’, MIT Press, pp. 841–848.
- O’Donoghue, B. & Candès, E. (2013), ‘Adaptive restart for accelerated gradient schemes’, *Foundations of Computational Mathematics* **15**(3), 715–732.
- Osher, S. & Li, Y. (2009), ‘Coordinate descent optimization for l1 minimization with application to compressed sensing; a greedy algorithm’, *Inverse Problems and Imaging* **3**(3), 487–503.
- Parikh, N. & Boyd, S. (2014), ‘Proximal algorithms’, *Foundations and Trends in Optimization* **1**(3), 127–239.
- Pearl, J. (2000), *Causality: models, reasoning, and inference*, Cambridge University Press.
- Pearl, J. (2009), ‘Causal inference in statistics: An overview’, *Statistics Surveys* **3**(0), 96–146.
- Pembrey, M. (2004), ‘The avon longitudinal study of parents and children (ALSPAC): a resource for genetic epidemiology’, *European Journal of Endocrinology* **151**, U125–U129.
- Ravikumar, P., Wainwright, M. J., Raskutti, G. & Yu, B. (2011), ‘High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence’, *Electronic Journal of Statistics* **5**, 935–980.
- Schmidt, M., Roux, N. L. & Bach, F. R. (2011), Convergence rates of inexact proximal-gradient methods for convex optimization, *in* ‘Advances in Neural Information Processing Systems 24’, Curran Associates, Inc., pp. 1458–1466.

- Shah, R. D. & Samworth, R. J. (2013), ‘Variable selection with error control: Another look at stability selection’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **75**(1), 5580.
- Sohn, K.-A. & Kim, S. (2012), Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization, in ‘Conference on Artificial Intelligence and Statistics’.
- Spirtes, P., Glymour, C. & Scheines, R. (2001), *Causation, Prediction, and Search*, The MIT Press.
- Stewart, G. & Sun, J.-g. (1990), *Matrix Perturbation Theory*, Academic Press.
- Stock, J. H. & Yogo, M. (2002), Testing for weak instruments in linear iv regression, Working Paper 284, National Bureau of Economic Research.
- Sutton, C. & McCallum, A. (2012), ‘An introduction to conditional random fields’, *FNT in Machine Learning* **4**(4), 267–373.
- The Economist (2010), ‘Data, data everywhere’, *The Economist* .
URL: <http://www.economist.com/node/15557443>
- Tütüncü, R. H., Toh, K. C. & Todd, M. J. (2003), ‘Solving semidefinite-quadratic-linear programs using sdpt3’, *Mathematical Programming, Series B* **95**(2), 189–217.
- Uhler, C., Raskutti, G., Bühlmann, P. & Yu, B. (2013), ‘Geometry of the faithfulness assumption in causal inference’, *Ann. Statist.* **41**(2), 436–463.
- Vandenberghe, L. & Boyd, S. (1996), ‘Semidefinite programming’, *SIAM Review* **38**(1), 4995.
- Wainwright, M. J. (2009), ‘Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso)’, *IEEE Transactions on Information Theory* **55**(5), 2183–2202.

- Wang, C., Sun, D. & Toh, K.-C. (2010), ‘Solving log-determinant optimization problems by a newton-cg primal proximal point algorithm’, *SIAM Journal on Optimization* **20**(6), 2994–3013.
- Wang et al. (2011), ‘NOA: a novel network ontology analysis method’, *Nucleic Acids Research* **39**(13).
- Westra, H.-J. et al. (2013), ‘Systematic identification of trans eQTLs as putative drivers of known disease associations’, *Nature Genetics* **45**(10), 1238–1243.
URL: <http://dx.doi.org/10.1038/ng.2756>
- Wytock, M. & Kolter, J. Z. (2013), Sparse Gaussian conditional random fields: Algorithms, theory, and application to energy forecasting, in ‘Proceedings of the 2013 International Conference on Machine Learning’, pp. 1265–1273.
- Ye, G.-B., Wang, Y. & Xie, X. (2011), ‘Efficient latent variable graphical model selection via split Bregman method’, *Available on the arXiv at <http://arxiv.org/pdf/1110.3076v1.pdf>*.
- Yin, J. & Li, H. (2011), ‘A sparse conditional Gaussian graphical model for analysis of genetical genomics data’, *Ann. Appl. Stat.* **5**(4), 2630–2650.
- Yuan, M. & Lin, Y. (2007), ‘Model selection and estimation in the gaussian graphical model’, *Biometrika* **94**(1), 19–35.
- Zhang, L. & Kim, S. (2014), ‘Learning gene networks under snp perturbations using eQTL datasets.’, *PLoS computational biology* pp. 1–20.