Graphical Model Selection for Gaussian Conditional Random Fields in the Presence of Latent Variables

Benjamin Frot, Luke Jostins & Gilean McVean

Graphical Model Selection for Gaussian Conditional Random Fields in the Presence of Latent Variables

Benjamin Frot[*]

Department of Statistics, University of Oxford

and

Luke Jostins

Wellcome Trust for Human Genetics, University of Oxford

and

Gilean McVean

Department of Statistics, University Of Oxford

January 16, 2018

Abstract

We consider the problem of learning a conditional Gaussian graphical model in the presence of latent variables. Building on recent advances in this field, we suggest a method that decomposes 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 estimator and show that it is well-behaved in the high-dimensional regime as well as "sparsistent" (*i.e.* capable of recovering the graph structure). We then show how proximal gradient algorithms and semi-definite programming techniques can be employed to fit the model to thousands of variables. Through extensive simulations, we

1

illustrate the conditions required for identifiability and show that there is a wide range of situations in which this model performs significantly better than its counterparts, for example, by accommodating more latent variables. Finally, the suggested method is applied to two datasets comprising individual level data on genetic variants and metabolites levels. We show our results replicate better than alternative approaches and show enriched biological signal.

2

1   Introduction

The task of performing graphical model selection arises in many applications in science and engineering. There are several factors that make this problem particularly challenging. First, it is common that only a subset of the relevant variables are observed and estimators that do not account for hidden variables are therefore prone to confounding. On the other hand, modelling latent variables is itself difficult because of identifiability and tractability issues. Second, the number of variables being modelled is often greater than the number of samples. It is well known that, in such a scaling regime, obtaining a consistent estimator is usually impossible without making further assumptions about the model, *e.g.* sparsity or low-dimensionality. Finally, modelling the joint distribution over all observed variables is not always relevant. It is sometimes preferable to learn a graphical model over a number of variables of interest while conditioning on the rest of the collection.

These problems are encountered in many fields of application. In genetics, for example, one might model a gene expression network conditional on the samples' combinations of DNA variants: the variables of interest are the expression levels, while the DNA variants are included because of their predictive power and capacity to explain some of the observed correlations between genes (Stearns 2010). As genotype is not causally influenced by gene expression levels (*i.e.* the direction of effect only goes genotype to expression), we would like to model expression levels conditional on genotype. For another example, consider the task of modelling stock returns conditional on sentiment analysis data. The variables that encode sentiment about the stocks have value (Li et al. 2014), but modelling their joint distribution might be difficult and unnecessary, hence the need for conditioning. Moreover, a number of unmeasured variables (*e.g.* energy prices) might impact many stocks and should be modelled for better predictive accuracy (Chandrasekaran, Parrilo & Willsky

3

2012).

The problem of learning a Gaussian graphical model in the presence of latent variables was considered by Chandrasekaran, Parrilo & Willsky (2012). They suggest estimating an inverse covariance matrix which is the sum of a sparse and a low-rank matrix. Another partial solution to our problem was introduced independently by Sohn & Kim (2012) and Wytock & Kolter (2013) who defined the concept of a *sparse Gaussian conditional random field*: a regularised maximum likelihood estimator that learns a Gaussian graphical model over a subset of the variables (X, say) while conditioning on the remaining variables (Z, say).

Chandrasekaran, Parrilo & Willsky (2012), Sohn & Kim (2012) and Wytock & Kolter (2013) made significant advances to the problem of model selection in general graphical models, but there exist many situations where we may wish to allow for latent variables and condition on some of those measured. Here, we suggest learning a Gaussian conditional random field in the presence of latent variables and introduce a novel regularised maximum likelihood estimator which fits into the "low-rank plus sparse" framework (Chandrasekaran et al. 2009, Candès et al. 2011). In our setting, inputs (variables in Z) are allowed to act on the outputs (X) in both a sparse and a low-rank fashion, while the inverse covariance matrix over X is estimated conditional on Z and on the marginalised latent variables. As will be shown later, this approach allows us to correctly recover graphs that are typically denser and with more hidden variables than the ones that can be handled by other methods.

From both a theoretical and a computational point of view, modelling latent variables with a conditional random field gives rise to a number of complications (*e.g.* the proximal operator is not defined in closed form) that we address in this paper. In particular, we derive convergence bounds for our estimator and show that under suitable identifiability conditions it is consistent in the high-dimensional regime as well as "sparsistent" (*i.e.* capable of

4

recovering the graph structure). We then show how the alternating direction method of multipliers (Boyd et al. 2010) and semi-definite programming techniques can be employed to fit the model to thousands of variables. Through extensive simulations, we illustrate the conditions required for identifiability and show that there is a wide range of situations in which this model performs significantly better than its counterparts. In order to show how our model behaves in a realistic setting, we apply the present estimator to two datasets comprising genetic variants and metabolite levels. Both replication and a test statistic constructed using an independent source of validation suggest that our estimates have more biological relevance than the results obtained via other methods.

2　Problem Statement

2.1　Setup

Throughout, we consider n independent, identically distributed realisations of a zero-mean random vector $Y \in \mathbb{R}^{m+p+h}$. Y is indexed by disjoint subsets of $\{1, \ldots, m+p+h\}$, 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). Our main assumption is that the distribution of $\begin{pmatrix} Y_X \\ Y_H \end{pmatrix} \in \mathbb{R}^{p+h}$ conditional on $Y_Z \in \mathbb{R}^m$ is normal and that its mean is a linear combination of the inputs Y_Z. Thus, conditioning on Y_Z, Y_X follows a multivariate normal distribution. More precisely, we assume a Gaussian conditional random field parametrised as follows:

$$\begin{pmatrix} Y_X \\ Y_H \end{pmatrix} | Y_Z \sim \mathcal{N} \left\{ - \begin{pmatrix} M_X^* & M_{XH}^* \\ M_{XH}^{*T} & M_H^* \end{pmatrix}^{-1} \begin{pmatrix} M_{ZX}^{*T} \\ M_{ZH}^{*T} \end{pmatrix} Y_Z, \begin{pmatrix} M_X^* & M_{XH}^* \\ M_{XH}^{*T} & M_H^* \end{pmatrix}^{-1} \right\},$$

5

where we have used partitioned matrices to show the contributions of the observed and hidden variables. Thus, $M_X^* \in \mathbb{R}^{p \times p}$, $M_{ZX}^* \in \mathbb{R}^{m \times p}$, $M_{XH}^* \in \mathbb{R}^{p \times h}$, ... The superscript $-^*$ is used to indicate that these matrices are parameters of the model, as opposed to estimates. Note that there are no distributional assumptions about Y_Z.

Finally, we assume that variables indexed by H are unobserved. Accordingly, we compute the marginal distribution $Y_X | Y_Z$, which yields

$$Y_X | Y_Z \sim \mathcal{N} \left\{ -\left(S_X^* - L_X^*\right)^{-1} \left(S_{ZX}^{*T} - L_{ZX}^{*T}\right) Y_Z, \left(S_X^* - L_X^*\right)^{-1} \right\}, \tag{2.1}$$

where we have defined $S_X^* \triangleq M_X^*$, $L_X^* \triangleq M_{XH}^* M_H^{*-1} M_{XH}^{*T}$, $S_{ZX}^* \triangleq M_{ZX}^*$ and $L_{ZX}^* \triangleq M_{ZH}^* M_H^{*-1} M_{XH}^{*T}$. This expression follows straightforwardly from the formula for the inverse of a partitioned matrix (the full derivation is given in the supplementary materials). From Equation (2.1), the log-likelihood function can be expressed in terms of the sample covariance matrices $\Sigma_Z^n \triangleq \frac{1}{n} \sum_i (Y_Z)_i (Y_Z)_i^T$, $\Sigma_X^n \triangleq \frac{1}{n} \sum_i (Y_X)_i (Y_X)_i^T$ and $\Sigma_{ZX}^n \triangleq \frac{1}{n} \sum_i (Y_Z)_i (Y_X)_i^T$:

$$\ell(S_X, L_X, S_{ZX}, L_{ZX}; \Sigma_Z^n, \Sigma_X^n, \Sigma_{ZX}^n) = \log \det \left(S_X - L_X\right) - Tr\left(\Sigma_X^n (S_X - L_X)\right) -$$
$$2Tr\left(\Sigma_{ZX}^n (S_{ZX} - L_{ZX})^T\right) - Tr\left(((S_X - L_X)^{-1}(S_{ZX} - L_{ZX})^T \Sigma_Z^n (S_{ZX} - L_{ZX})\right). \tag{2.2}$$

For clarity, all terms related to a given subset will be dropped from the expression when the subset is empty. For example, whenever $Z = H = \emptyset$ the log-likelihood becomes $\ell(S_X; \Sigma_X) = \log \det S_X - Tr(\Sigma_X^n S_X)$.

Note that our assumption about the Gaussianity of X, H plays an important role in the interpretation of the parameters $(M_X^*, M_{XH}^*, \ldots)$. Under this assumption, it is well known that the structure of the conditional Gaussian graphical model (GGM) over X, H can be read-off these matrices directly by looking at the location of their non-zero entries (Lauritzen 1996). Briefly, a graphical model is a statistical model defined according

6

to a graph whose nodes are random variables and whose edges encode conditional independence statements between variables (Lauritzen 1996). Thus, $(M_X^*)_{i,j} = (M_X^*)_{j,i} = 0$ if and only if $X_i \perp\!\!\!\perp X_j | Z, X \setminus \{X_i, X_j\}, H$. Likewise, $(M_{ZH}^*)_{i,j} = 0$ if and only if $Z_i \perp\!\!\!\perp H_j | Z \setminus Z_i, X, H \setminus \{H_j\}$. Note that since the conditional mean vector is a linear transformation of Y_Z, this interpretation of the non-zero entries of M_{ZH}^* and M_{ZX}^* holds irrespective of Y_Z's distribution.

2.2 Goal

In typical applications such as the ones mentioned in the introduction, S_X^* is the target. Since it encodes the structure of the graphical model over X, recovering S_X^* can provide insight into the causal mechanisms underpinning the data but, in general, hidden variables make it impossible to access this parameter directly. Instead, it follows from Equations (2.1) and (2.2) that only the *sum* $S_X^* - L_X^*$ can be inferred (similarly, only $S_{ZX}^* - L_{ZX}^*$ is accessible). The maximiser of the log-likelihood (2.2) is not unique and the problem is fundamentally misspecified.

We are therefore facing two related, but distinct, problems:

- *identifiability*: under which conditions does the problem admit a *unique* solution? Ideally, these conditions ought to be as broad as possible so that they will be met in realistic situations. Note 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, can we derive a consistent, tractable estimator which is capable of recovering $(S_X^*, L_X^*, S_{ZX}^*, L_{ZX}^*)$?

7

Here, we chose to focus on S_X^* because it fits our application but there might be situations in which other parameters are of interest, *e.g.* S_{ZX}^* in Zhang & Kim (2014).

2.3 Previous Work

In practice, model selection in the context of GGMs is often performed using ℓ_1-regularised maximum likelihood estimators (MLEs) such as the ones introduced in Banerjee et al. (2008), Yuan & Lin (2007), and the so-called *graphical lasso* (Friedman et al. 2008). 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). Beyond their attractive computational properties, the ℓ_1 and nuclear norm regularised MLEs enjoy strong theoretical guarantees (Bach 2008, Ravikumar et al. 2011).

Using penalised MLEs, the questions raised above (Section 2.2) have been solved in some special cases of model (2.1).

Sparse Gaussian Conditional Markov Random Field: $H = \emptyset$

When H is empty, (2.1) reduces to

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

The log-likelihood associated with this model is convex and maximum-likelihood estimates can be obtained in closed form. In order to increase the interpretability of the estimates

8

and cope with high-dimensionality, Sohn & Kim (2012), Wytock & Kolter (2013) suggested the following estimator of (S_X^*, S_{ZX}^*):

$$(\hat{S}_X, \hat{S}_{ZX}) = \underset{S_X \in \mathbb{R}^{p \times p}, S_{ZX} \in \mathbb{R}^{m \times p}, S_X \succ 0}{\arg\min} -\ell(S_X, S_{ZX}; \Sigma_Z^n, \Sigma_X^n) + \lambda_n(||S_X||_1 + ||S_{ZX}||_1),$$

with $\lambda_n > 0$. 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 S_X^* and S_{ZX}^*. Among other results, they showed that this approach often outperforms the graphical lasso in terms of predictive power and model selection accuracy. Alternative parametrisations and approaches have been suggested in the multivariate linear regression literature. We refer the reader to Yin & Li (2011), Sohn & Kim (2012) and references therein for more details on these estimators and their relative performances.

Low-Rank Plus Sparse Decomposition: $Z = \emptyset$

The presence of latent variables $(H \neq \emptyset)$ is a substantial complication. As explained earlier, the marginal precision $S_X^* - L_X^*$ is then the sum of two matrices and 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 $S_X^* - L_X^*$ into its summands. Loosely speaking, this is the case if S_X^* is sparse and there are few hidden variables with an effect spread over most of the observed variables. As a result, Chandrasekaran, Parrilo & Willsky (2012) introduced an estimator which penalises the ℓ_1-norm of S_X and the nuclear norm of L_X as follows:

$$(\hat{S}_X, \hat{L}_X) = \underset{S_X, L_X \in \mathbb{R}^{p \times p}}{\arg\min} -\ell(S_X, L_X; \Sigma_X) + \lambda_n(\gamma||S_X||_1 + ||L_X||_*), \tag{2.3}$$

9

subject to $S_X - L_X \succ 0, L_X \succeq 0.$, with $\lambda_n, \gamma > 0$. Here $||L_X||_*$ denotes the nuclear norm of L_X (*i.e.* the sum of its singular values). 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.

2.4   Suggested Estimator

As hinted in the introduction, there are many cases where one might want to both condition and allow for latent variables. In such cases, neither the sparse Gaussian conditional Markov random field nor the low-rank plus sparse approach would be optimal. Building on these estimators, we propose decomposing the parameters of a Gaussian conditional Markov random field into the sum of a low-rank and a sparse matrix. To that end, we suggest optimising the following regularised MLE

$$(\hat{S}_X, \hat{L}_X, \hat{S}_{ZX}, \hat{L}_{ZX}) =$$

$$\underset{S_X, L_X \in \mathbb{R}^{p \times p}; S_{ZX}, L_{ZX} \in \mathbb{R}^{m \times p}}{\arg \min} -\ell(S_X, L_X, S_{ZX}, L_{ZX}; \Sigma_Z^n, \Sigma_X^n, \Sigma_{ZX}^n) + \lambda_n(\gamma \|S\|_1 + \|L\|_*)$$

$$\text{s.t. } S_X - L_X \succ 0, L_X \succeq 0 \text{ and } S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}. \quad (2.4)$$

Solving (2.4) amounts to minimising a function which is *jointly convex* in its parameters over a convex constraint set (proofs are in the supplementary materials, along with other elementary properties of the likelihood). As mentioned earlier, this likelihood is structured around two parameters, S_{ZX} and S_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 nuclear norm of L, the effect of all latent variables is modelled jointly and

10

a single set of latent factors is learned. No distinction 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.

3   Theoretical Analysis

According to our assumptions, we assume here that each sample is generated according to the model

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

and ask under what circumstances Estimator (2.4) correctly recovers the parameters S^*, L^* (as built by stacking S_X^*, S_{ZX}^* and L_X^*, L_{ZX}^*) with overwhelming probability.

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.

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: they are expressed in terms of the Fisher information matrix but do not explicitly involve the likelihood function. For that reason, key definitions, as well as assumptions necessary for our result to hold, are deferred to the supplementary materials. Here we focus on the original contributions of this paper by giving an intuition for these conditions before formally stating the *consistency* of the estimator defined by (2.4).

11

3.1 Identifiability

Until now, it was repeatedly mentioned that a "low-rank plus sparse decomposition" is possible when 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 a 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 the conditions under which such a problem is well-posed, even for arbitrary matrices. Two concepts are particularly important (precise mathematical statements and explanations can be found in the supplementary materials):

- $\xi(T(L^*))$: a small $\xi(T(L^*))$ guarantees that no single latent variable will have a strong effect on only a small set of the observed variables. It is closely related to the concept of *incoherence* found in Candès et al. (2011).

- $\mu(\Omega(S^*))$ quantifies the diffusivity of S's spectrum. It can be shown that matrices with few non-zero entries per row/column (and thus sparse) have a small μ.

A sufficient condition for identifiability can be expressed in terms of ξ, μ by requiring that their product be small enough ($(\xi(T(L^*))\mu(\Omega(S^*)) \leq \frac{1}{6}\mathcal{C}^2$) and that the tuning parameter γ be chosen within a given range ($\gamma \in \left[\frac{3\xi(T(L^*))}{\mathcal{C}}, \frac{\mathcal{C}}{2\mu(\Omega(S^*))}\right]$), for some constant $\mathcal{C}$ which depends on the Fisher information matrix (FIM). In other words, there must be a small number of latent variables acting on many observed ones and S^* must not have too many non-zero entries in any given row or column. This is a condition on the parameters S^*, L^* and it is related to the problem of decomposing the sum of two matrices. Moreover, it can be shown that natural classes of matrices satisfy these assumptions. In particular, the degree of S^* (q) and number of latent variables (h) are allowed to grow as a function of

12

the problem size p, m (Chandrasekaran et al. 2009). For example, under some assumptions about the distribution from which L^* is sampled, one shows that $\xi(T(L^*)) \sim \sqrt{\frac{h}{p}}$ and that scaling regimes of the form $q \sim \log(p+m)^b$ and $h \sim \frac{p}{\log(p+m)^{2b}}$ (for $0 \leq b < \infty$) guarantee identifiability with high probability (see §4.2 in Chandrasekaran, Parrilo & Willsky (2012)). We call these restrictions on ξ, μ and γ **Assumption 1**.

Another issue is that one does not directly observe $S^* - L^*$ but samples generated from (3.1). All lasso-type methods face this problem and conditions on the FIM are usually imposed (the so-called *irrepresentability condition*) (Ravikumar et al. 2011). Similar assumptions about the FIM are made here and detailed in the supplementary materials. This is **Assumption 2**.

3.2 Consistency

We can now present our main result and state the consistency of Estimator (2.4) (see supplementary materials for the proof). First, let us recall that for any matrix P, $||P||_2$ denotes its largest singular value and $||P||_\infty$ is its largest entry in magnitude. We can then define the following quantities:

$$\psi_Z = ||\Sigma_Z^n||_2, \ \psi_X^* = ||(S_X^* - L_X^*)^{-1}||_2, \ \phi_{ZX}^* = ||S_{ZX}^* - L_{ZX}^*||_2,$$

$$\psi = 2\psi_X^* \sqrt{1 + 6\frac{\psi_Z}{\psi_X^*}(1 + \psi_X^* \phi_{ZX}^*)^2}, W = Q_1 \min\left(\frac{1}{4\psi_X^*} \frac{Q_2}{\psi_X^* \psi^2}\right).$$

Finally, for $M = \max\left(1, \frac{\psi_Z}{4\psi_X^*}(1 + \sqrt{\frac{m}{p}})^2\right)$, let $\lambda_n = \frac{Q_3}{\xi(T(L^*))}\sqrt{\frac{256\psi_X^{*2}pM}{n}}$.

We prove the following theorem in the supplementary materials (Q_1 to Q_6 are constants whose definitions are deferred for clarity):

13

Theorem 1. *(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:

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

(b) $(\sigma_{\min}$ *and* $\theta_{\min}$ *conditions) Let the minimum non-zero singular value* σ *of* L^* *and the minimum non-zero entry of* S^* *in magnitude* θ *be such that*

$$\sigma \geq \frac{Q_4 \lambda_n}{\xi(T(L^*))^2}, \ \ \theta \geq \frac{Q_5 \lambda_n}{\mu(\Omega(S^*))}.$$

Then, with probability greater than $1 - \max\left(2\exp(-pM), \exp(-4\frac{\psi_X^*}{\varphi_2}pM)\right)$ *we have*

1. $sign(\hat{S}) = sign(S^*)$, $rank(\hat{L}) = rank(L^*)$ *and* $\hat{L}_X \succeq 0$.

2.
$$\max\left(\frac{1}{\gamma}\left\|\hat{S} - S^*\right\|_\infty, \left\|\hat{L} - \bar{L}^*\right\|_2\right) \leq \frac{Q_6 \psi_X^*}{\xi(T(L^*))}\sqrt{\frac{pM}{n}}.$$

Seen at a high-level, Theorem 1 is analogous to the result obtained by Chandrasekaran, Parrilo & Willsky (2012) for the low-rank plus sparse (LR+S) estimator. A particularly important feature is that **Assumption 1** holds even the degree of S^* and the rank of L^* grow with the dimensions of the problem. Taking as example the scaling regime mentioned in §3.1, we see that $n \gtrsim p\log(p+m)^{4b}M$ samples are required for Th. 1 to hold with high probability. This is also enough to guarantee that $\max\left(\frac{1}{\gamma}\left\|\hat{S} - S^*\right\|_\infty, \left\|\hat{L} - L^*\right\|_2\right) = o_P(1)$ but it should be contrasted with the logarithmic scaling usually encountered in the ℓ_1-regularised literature, *e.g* one asks $\frac{\log(pm)}{n} = o(1)$ for the SCGGM estimator of Wytock & Kolter (2013).

In order to further compare the convergence rates of our estimator and LR+S, a few points are worth considering.

14

First, we do not make any distributional assumptions about Y_Z and there are therefore many scenarios in which only Th. 1 applies. For the sake of comparison, we can assume that Y_Z follows a normal distribution so that the consistency theorem of LR+S is applicable. Since LR+S does not model a conditional distribution, Z and X are modelled jointly. The estimated matrices, $(\hat{S}_{LR+S}, \hat{L}_{LR+S})$, are of size $(p+m) \times (p+m)$ and, to obtain $\hat{S}_X, \hat{S}_{ZX}, \ldots$, the relevant sub-matrices are extracted from the larger $(p+m) \times (p+m)$ estimates. Considering the scaling described in §3.1, we see that high-dimensional regimes of the form $p_n + m_n = \mathcal{O}(n^{1-a})$ for $0 < a \leq 1$ cover interesting applications and are enough to guarantee consistency. To see "why our convergence rates are similar" (CHANGE!), start by noticing that under the **Assumption 1** of either theorem $m_n = o(n)$ is required for consistent estimation. Now, since Y_Z follows a normal distribution, we have that $\lim_{n\to\infty} \psi_Z = \sigma^2(1+\sqrt{y})^2$, for some $0 \leq \sigma < \infty$ and where $y = \lim_{n\to\infty} m_n/n$ (see *e.g.* Th. 5.11 in Bai & Silverstein (2009)). Therefore, $\psi_Z = \mathcal{O}(1)$ and $p_n M_n = \mathcal{O}((\sqrt{p_n} + \sqrt{m_n})^2) = \mathcal{O}(p_n + m_n)$ so that assuming $p_n + m_n = \mathcal{O}(n^{1-a})$ for $0 < a \leq 1$ is also enough for consistent estimation in many settings of interest.

Second, μ and ξ play an identical role in both Theorem 1 and (Chandrasekaran, Parrilo & Willsky 2012, Theorem 4.1), namely through **Assumption 1** and conditions a) and b). However, these quantities are usually different (*i.e.* $\mu(\Omega(S^*)) \neq \mu(\Omega(S^*_{LR+S}))$, $\xi(T(L^*)) \neq \xi(T(L^*_{LR+S}))$), which has interesting implications. An obvious consequence is the one stated in the previous section: since μ, ξ define the acceptable range for γ, its span can vary widely across methods. More importantly, one shows that conditions a) and b) are driven by the lower-end of that range. Should it be assumed instead that $\gamma = \frac{\mathcal{C}}{2\mu(\Omega(S^*))}$ (the upper-end), all three conditions would be relaxed (Chandrasekaran, Parrilo & Willsky 2012, Corollary 4.2). Thus, the smaller the value of $\xi(T(L^*))$, the wider the acceptable range and the more likely Theorem 1 is to hold.

15

4 Optimisation

Optimising (2.4) in the high-dimensional setting is a challenging problem. For example, some of the constraints are hard to accommodate (*e.g.* $S_X - L_X \succ 0, L_X \succeq 0$) and the penalty terms are non-smooth. Fortunately, (2.4) has similarities with (2.3) (the estimator of Chandrasekaran, Parrilo & Willsky (2012)) and we can rely on algorithms that have proven effective on (2.3), namely the Alternative Direction Method of Multipliers (ADMM) (Boyd et al. 2010, Ma et al. 2013, Ye et al. 2011) and approaches relying on Semi-Definite Programming (SDP) (Vandenberghe & Boyd 1996, Wang et al. 2010, Tütüncü et al. 2003). The general theory behind both ADMM and SDP is applicable to the problem at hand but features that are specific to (2.4) prevent a straightforward application of existing algorithms. SDP is an active field of research and recasting (2.4) within that framework makes it easier for the reader to use existing software and even benefit from future advances in that field. On the other hand, our ADMM implementation is tailored to the problem at hand but converges to a reasonable accuracy quickly. This is why we discuss both strategies. Technical details and step-by-step derivations are given in the supplementary materials.

4.1 The Alternating Direction Method of Multipliers

The Alternating Direction Method of Multipliers (ADMM) is a first-order optimisation procedure which is well-suited to the minimisation of large-scale convex functions. It proceeds by decomposing the original problem into more amenable subproblems which are then solved iteratively (Boyd et al. 2010). It is sometimes possible to obtain closed-form solutions to these subproblems but this is not required for ADMM to converge: even inexact iterative methods can be employed (Eckstein & Bertsekas 1992, Goldstein & Osher

16

2009). Moreover, only a few tens of iterations are necessary for ADMM to converge to an accuracy which is sufficient for most applications[1] (Boyd et al. 2010). ADMM (and related algorithms such as Bregman iterations and Douglas–Rachford splitting) has been celebrated as an efficient and robust general-purpose algorithm for ℓ_1-regularised problems (Goldstein & Osher 2009).

More recently, (Ye et al. 2011, Ma et al. 2013) used ADMM to solve (2.3) and showed that it can be optimised by iteratively solving four smaller subproblems (Ye et al. 2011). A similar decomposition is applicable to the problem at hand but, in the case of (2.4), one of the subproblems requires the computation of a so-called *proximal operator* which does not admit a closed-form solution. Consequently, we derived an algorithm which iteratively converges to this proximal operator. In practice, we found that only a few iterations (typically less than 10) of this subprocedure are necessary to obtain a good approximation to the proximal operator.

4.2 Recasting the objective function as a Semi-Definite Program

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):

$$\underset{X_1, X_2, \dots}{\arg\min} \ Tr(X_1 C_1^T) + Tr(X_2 C_2^T) + \dots + a_1 \log \det(X_1) \tag{4.1}$$

subject to a number of linear, quadratic and positive semi-definite constraints [2]. Our goal is then to recast (2.4) as a problem of the same form as (4.1). We show in the supplementary

---

[1]However, converging to a very high accuracy can be slow in comparison to second-order methods.

[2]This is only a subset of the problems that can be tackled by such packages. See references for a formulation of this problem in its full generality.

17

materials that (2.4) admits the following SDP reformulation:

$$\underset{S_X,L_X,S_{ZX},L_{ZX},W,F,H_1,H_2}{\arg\min} Tr(K\Sigma_O^n) - \log\det S_X + \lambda_n\left(\gamma\mathbf{1}^T F\mathbf{1} + \frac{1}{2}\left(Tr(H_1)+Tr(H_2)\right)\right)$$

$$\text{subject to } K \succeq 0, \ S_X \succ 0, \ L_X \succeq 0, \ \begin{pmatrix} H_1 & L \\ L^T & H_2 \end{pmatrix} \succeq 0, \ -F_{ij} \leq S_{ij} \leq F_{ij}, \ \forall i,j;$$

$$\text{where } K = \begin{pmatrix} W & S_{ZX}-L_{ZX} \\ S_{ZX}^T-L_{ZX}^T & S_X-L_X \end{pmatrix}, \ S = \begin{pmatrix} S_X \\ S_{ZX} \end{pmatrix}, \ L = \begin{pmatrix} L_X \\ L_{ZX} \end{pmatrix}. \quad (4.2)$$

(4.2) can easily be implemented in *e.g.* YALMIP and solved using LogdetPPA or SDPT3 (Löfberg 2004, Wang et al. 2010, Tütüncü et al. 2003). We remark that the objective function is now smooth (as opposed to (2.4)) but contains many more variables and constraints.

5 Simulations

We now study the properties of the proposed model on synthetic data and compare its performances to the three other methods introduced earlier: the graphical lasso (GLASSO) (Friedman et al. 2008), the sparse conditional Gaussian graphical model (SCGGM) (Sohn & Kim 2012, Zhang & Kim 2014, Wytock & Kolter 2013) and the low-rank plus sparse decomposition (LR+S) (Chandrasekaran, Parrilo & Willsky 2012). The suggested approach will henceforth be referred to as *LSCGGM* (*i.e.* Latent Sparse Conditional Gaussian Graphical Model).

In Section 3, it was established that assumptions about both the nominal parameters (S^*, L^*) and the Fisher information matrix are necessary to guarantee the identifiability of the problem and, subsequently, the applicability of Theorem 1. In particular, we recalled the key role played by the maximum degree of S^* and the incoherence of L^*. To better

18

understand when these assumptions are expected to hold, we simulate data from a set of graphical models that span the range of possible latent structures and measure the ability of the different methods to recover the underlying graphs.

5.1   Graphical Structures and Methods

The set of graphical structures we simulate from is constructed in such a way that only two integers, d_Z and d_H, describe the relevant properties (rank, sparsity, incoherence, degree) of $S_{ZX}^* - L_{ZX}^*$ and L_X^*, respectively. Thus, d_Z controls the relationship between inputs (Z) and outputs (X) while d_H encodes the behaviour of L_X^*. The remaining parameter, S_X^*, remains unchanged throughout. We now briefly describe how the graphs are constructed but defer technical details to the supplementary materials (*e.g.* distribution of effect sizes). The code used to generate the data and fit our model is made available with this paper.

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

$$\begin{pmatrix} Y_X \\ Y_H \end{pmatrix} | Y_Z \sim \mathcal{N} \left\{ - \begin{pmatrix} S_X^* & M_{XH}^* \\ M_{XH}^{*T} & M_H^* \end{pmatrix}^{-1} \begin{pmatrix} M_{ZX}^{*T} \\ 0 \end{pmatrix} Y_Z, \begin{pmatrix} S_X^* & M_{XH}^* \\ M_{XH}^{*T} & M_H^* \end{pmatrix}^{-1} \right\},$$

with Y_Z a random vector of size p whose entries are drawn independently from a t-distribution with 4 degrees of freedom. Y_X is also of size p. Here, Y_X and Y_H are drawn jointly from a conditional random Markov field but only Y_X and Y_Z are observed, which implies that $L_X^* = M_{XH}^* M_H^{*-1} M_{XH}^{*T}$. The matrices S_X^*, L_X^* and M_{ZX}^* are constructed as follows.

The non-zero pattern of the $p \times p$ matrix S_X^* is identical across all simulations and is similar to the one adopted by Wytock & Kolter (2013): the graph over X is a chain of p variables in which one link out of five has been removed. The non-diagonal entries of S_X^* are such that $S_{Xij}^* \neq 0$, if and only if $i = j + 1$ and $i \not\equiv 0 \pmod 5$.

19

As stated above, the rank/sparsity of L_X^* is described by a single integer, d_H. Specifically, we assume that p is an integer of the form $p = 2^k$ and pick $d_H \in \{0, 1, \ldots, k\}$. Then, for a fixed value of d_H, M_H^* and M_{ZX}^* are random matrices constructed so that: a) there are exactly 2^{d_H} confounders, *i.e.* the rank of L_X^* is 2^{d_H}; b) each of the 2^{d_H} confounders impacts exactly $p/2^{d_H}$ outputs; c) each output is connected to exactly one latent variable. Thus, when $d_H = k$, there is effectively no confounding since latent variables and outputs are in a one-to-one correspondence. When d_H is much smaller than k, there are few confounders with an effect spread over many observed variables. When d_H is set close to k, there are many hidden variables, each affecting a handful of outputs – a gross violation of the identifiability assumptions.

Likewise, d_Z accounts for the structure of M_{ZX}^*. Here again, we assume $p = 2^k$ and pick $d_Z \in \{0, 1, \ldots, k\}$. Then, M_{ZX}^* is designed to satisfy: a) $rk(M_{ZX}^*) = 2^{d_Z}$; b) each row/column of M_{ZX}^* has exactly $p/2_Z^d$ non-zero entries. The effect of d_Z is easily interpreted. For example, $d_Z = k$ is an ideal situation where inputs and outputs are in a one-to-one correspondence. As d_Z goes from k to 0, M_{ZX}^* becomes denser and increasingly incoherent. When d_Z is close to k, M_{ZX}^* is estimated as a sparse matrix. When d_Z is small, its decomposition is a single low-rank matrix.

Finally, since neither GLASSO nor LR+S model conditional distributions, we use these estimators as described in Section 3, *i.e.* by first modelling Z and X jointly and then extracting submatrices of the estimates.

5.2   Results

In our simulations, we set $p = 32$, $n = 3000$ and let (d_Z, d_H) take values in $\{2, 3, 4, 5\}^2$. Each of these 16 designs is replicated 20 times, for a total of 320 distinct datasets.

Here, we are interested in recovering the structure of S_X^* and we use precision/recall

20

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 distinct precision/recall curve by varying λ. For each of the 320 simulated datasets, we computed the paths corresponding to 15 distinct values of γ and subsequently selected γ so as to maximise the Area Under the Curve (AUC). Figure 1 shows the average precision/recall curves obtained by applying this procedure.

First, we see that known methods behave as expected: GLASSO behaves best when there is no confounding and Z acts in a sparse fashion ($d_H = d_Z = 5$); SCGGM is more robust to changes in d_Z, but this is restricted to situations in which there is not confounding ($d_H = 5$); LR+S performs best when $d_H = 5$ or when there is low-rank, diffuse confounding ($d_H = 2$). In a number of cases, the method proposed here is better than any of the alternative methods and, in the worst cases, it offers comparable performances. Specifically, it outperforms LR+S significantly when both inputs and hidden variables act on the outputs through a relatively low-rank mechanism ($d_Z = 2, 3; d_H = 2, 3$). Two factors might explain this behaviour: a) the inputs are not normally distributed, which violates the assumptions of LR+S; b) the data is generated according to a conditional random Markov field, which is not assumed by LR+S, and may result in a violation of its identifiability assumptions.

$d_H = 4$ corresponds to the extreme situation in which each latent variable confounds exactly two random variables. None of the methods performs well but LR+S and LSCGGM behave better than GLASSO and SCGGM in scenarios where one would not expect any differences (*e.g.* $d_Z = d_H = 5$). This is because LR+S and LSCGGM have *two* tuning parameters, one of which (γ) is chosen with perfect knowledge: it improves the AUC of these methods but causes $\hat{L}_X$ to be non-zero. Additional simulations made available in the supplementary materials show that when γ is chosen with cross-validation, the selected
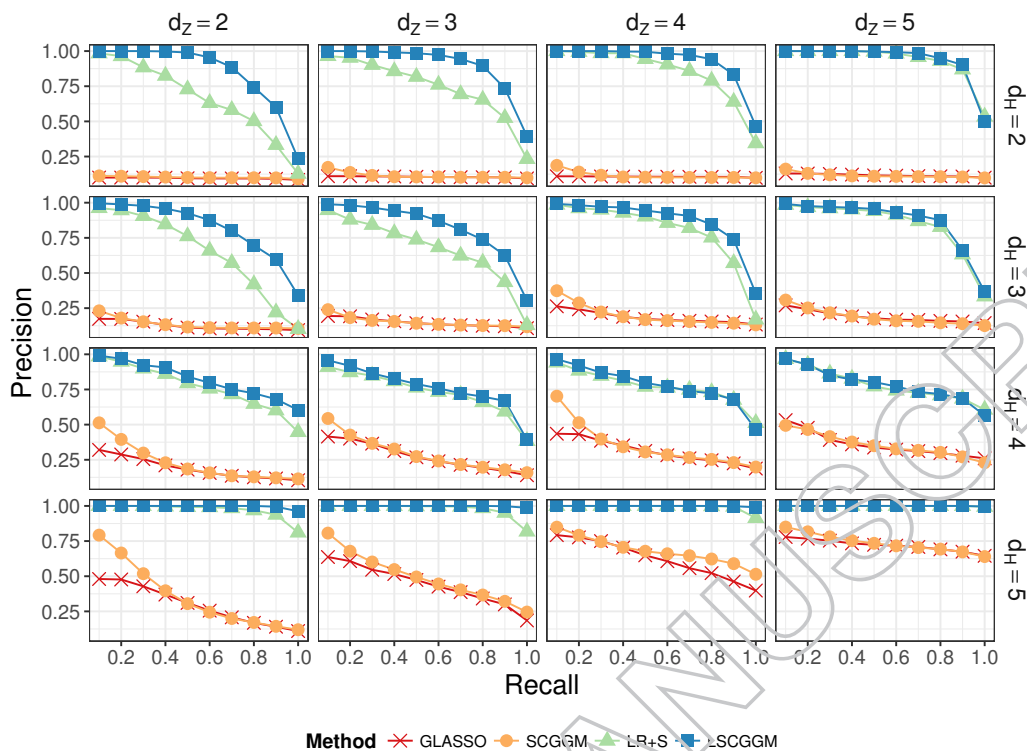
21

Figure 1: Comparison of the suggested estimator (LSCGGM) to other published methods. Along the x-axis (resp y-axis), d_Z (resp d_H) varies from 2 to 5. More precisely, in the bottom row ($d_H = 5$), there is no confounding at all. In the second row from the bottom ($d_H = 4$), hidden variables act in a very sparse fashion. In the top row ($d_H = 2$), there are 4 hidden variables and we are in the range of applicability of the low-rank plus sparse method. The second row ($d_H = 3$) corresponds to an intermediate regime in which there are 8 latent variables. Settings: $p = 32$, $n = 3000$. For each dataset, the value of the tuning parameter γ was chosen so as to maximise the Area Under the Curve (AUC). Each of the 16 designs is repeated 20 times. We report average precisions at fixed recalls of $\{0.1, 0.2, \ldots, 1\}$.

22

value of γ is indeed often too small.

Both LR+S and LSCGGM have two tuning parameters (λ, γ): λ controls the overall shrinkage on the sparsity/rank of the estimates, γ accounts for the trade-off between sparse and low-rank components. To better understand the role of γ, we look at the precision/recall curves obtained for various values of this tuning parameter. As suggested in Chandrasekaran et al. (2009), the penalty term is reformulated as $\lambda(\gamma||S||_1 + (1-\gamma)||L||_*)$ with γ ranging from 0 to 1 instead of $(0, +\infty)$. By analogy to the AUC metric, we report the "Volume Under the Surface" (VUS) which accounts for the effect of both regularisation parameters.

In Figure 2, the surfaces obtained for $(d_H = d_Z = 2)$ and $(d_H = 5, d_Z = 3)$ are plotted. They show that the suggested approach is less sensitive to γ than LR+S, thus making it easier to pick a sensible value in real-world applications. Figure 2 b) illustrates what happens when both methods offer comparable performances according to Figure 1 (which is obtained by choosing γ *perfectly*): compared to LSCGGM, there are actually very few values of γ for which LR+S achieves its best AUC. Here, only two of the 16 possible surface plots are shown, but Figure 2 c) indicates that LSCGGM is less sensitive to this tuning parameter across all simulation designs, as measured by the VUS. In particular, we have consistently observed that upper-end of the acceptable range for γ is higher for LSCGGM than LR+S. The next simulations illustrate the implications of this property.

In these simulations, our main concern was to illustrate how methods differ in terms of identifiability and consistency. Setting p and m to a relatively small value (32) made it possible to capture most scenarios with only 16 graphical structures. In the supplementary materials, we simulate from larger graphs ($p = m = 2^7 = 128$, $n = 3000$) and obtain results that are similar to the ones showed here. We also report the estimation errors for the other L^* along with the precision/recall curves for S_{ZX}^*. Finally, we look at the effect of choosing

23

γ using cross-validation. In the next section, we show how one can select λ and γ when some control over the number of falsely discovered edges is expected.
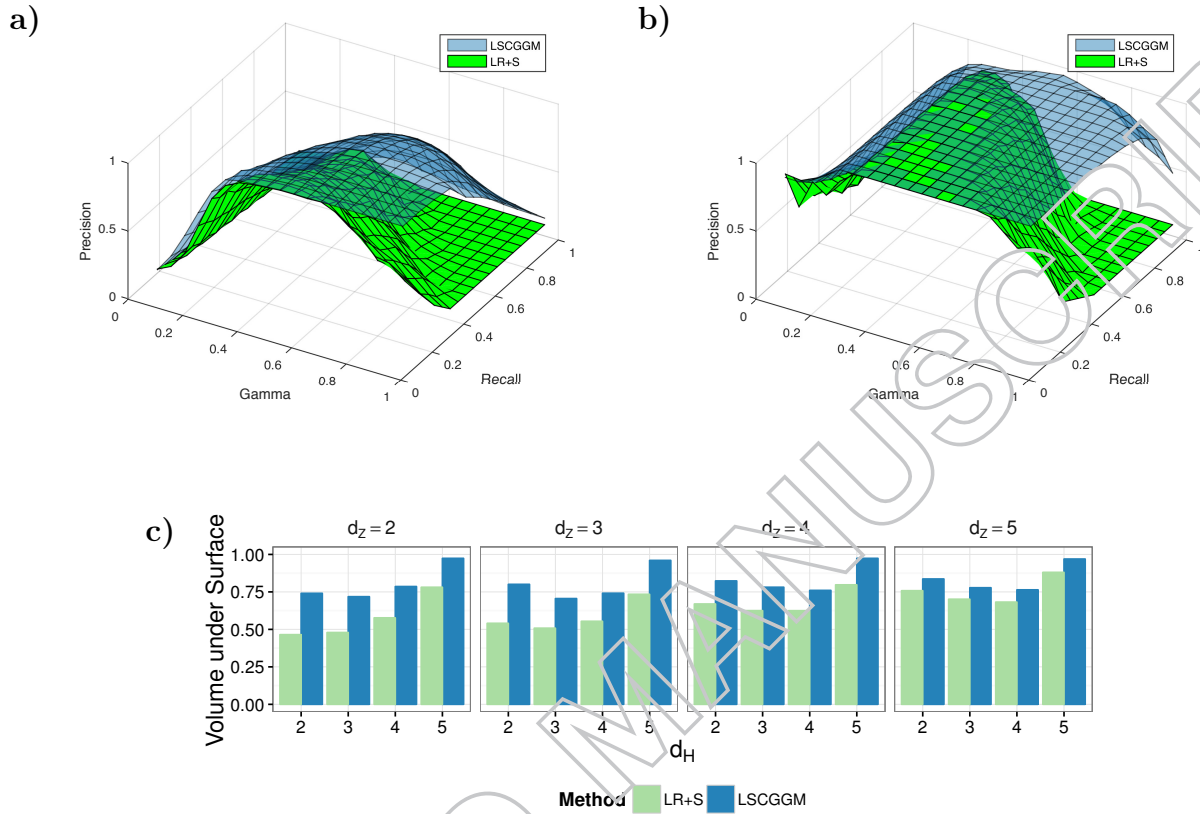

Figure 2: Sensitivity to the tuning parameter γ. Here, an alternative parametrisation of the regularisation term is used: $\lambda_n(\gamma||S||_1 + (1-\gamma)||L||_*)$, so that $\gamma \in (0,1)$ instead of $(0,+\infty)$. a) Precision/recall surface for $d_Z = d_H = 2$ (*i.e.* each input acts on 8 random outputs and there are 4 confounding variables). b) Precision/recall surface for $d_Z = 3$ and $d_H = 5$ (there are no confounders, each input acts on 4 random outputs). c) Volume under surface across all 16 simulation designs.

24

6 Application: Using genetic information to detect relationships between human metabolites

To illustrate the value of our new approach, we now apply it 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 the inputs, Z. For comparison purposes, we also report the results obtained with the Low-Rank plus Sparse method (LR+S)[3].

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 (Boyd et al. 2012, Fraser et al. 2012). More details about this study and data preparation are available in the supplementary materials. Here, only key features of this dataset are reported[4].

The data at our disposal contains genetic and phenotypic measurements on approximately 8,000 children and their mothers. We first performed our entire analysis on the children's cohort (called "Child cohort" throughout) and then independently applied the same procedure to the mothers' cohort (Mother cohort). We modelled the levels of 39 metabolites. Measurements for all 39 variables were available without missing data for 5,242 children and 2,770 mothers. In each cohort, independent genetic variants were se-

---

[3]The other two methods (graphical lasso and sparse conditional graphical model) arise as special cases by setting γ close to 0. For completeness, the results obtained by applying SCGGM are reported in the supplementary materials.

[4]Please 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.

lected based on their predictive power with respect to any of the 39 traits under study: 133 and 44 variants were selected 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 $(0/1/2)$.

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

6.2 Methods

Since both the suggested approach (LSCGGM) and the LR+S method have two tuning parameters (λ, γ), some procedure is required in order to set these parameters to appropriate values. As shown by both theoretical results and simulations, solutions are expected to be identical for a range of values of γ. Consequently, we do not select a single value of γ but consider instead 30 values within the range $(0.02, 0.98)^5$. To each γ corresponds a regularisation path: a graph along each path is selected using "pointwise" complementary pairs stability selection (Meinshausen & Bühlmann 2010, Shah & Samworth 2013). 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 their notations). Thus, for each method and each cohort we obtain a collection of thirty graphical structures.

In order to measure how similar two graphical structures are, we consider 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

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

---

[5] The penalty is parametrised as $\lambda(\gamma ||S||_1 + (1 - \gamma)||L||_*)$, so that $\gamma \in (0, 1)$.

26

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.

Another important step is assessing the biological relevance of the estimates using an external source of information. We used ChEBI: an ontology of small chemical entities of biological interest (Hastings et al. 2012). We manually matched all 39 metabolites to their ChEBI IDs and annotated them using the ontology. Using such annotations, one can compute an "enrichment statistic" reflecting whether a given graph contains edges between related metabolites more often than would be expected in a random graph with a similar topology (such a graph has an expected statistic of 1). We defer the definition of this statistic to the supplementary materials but remark that this method is close to the ontology analyses frequently encountered in computational biology (Wang et al. 2011). By randomly permuting the annotations, empirical p-values for this statistic can also be computed.

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 3a 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 $0.6 \le \gamma_1, \gamma_2 \le 0.9$, they are very close to each other. For small values of γ, the graphical structures returned by LSCGGM vary smoothly with γ. The regime $\gamma \le 0.05$ corresponds to the case in which the rank of the latent component is 0: LSCGGM behaves

27

like a sparse conditional graphical model. Similar figures can be found for the LR+S method and the Mother cohort in the supplementary materials.

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) (Figure 3b). For small values of γ, LR+S and LSCGGM appear 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. Figure 4 shows in what way the graphs found in those stable regions differ, with LR+S inferring more
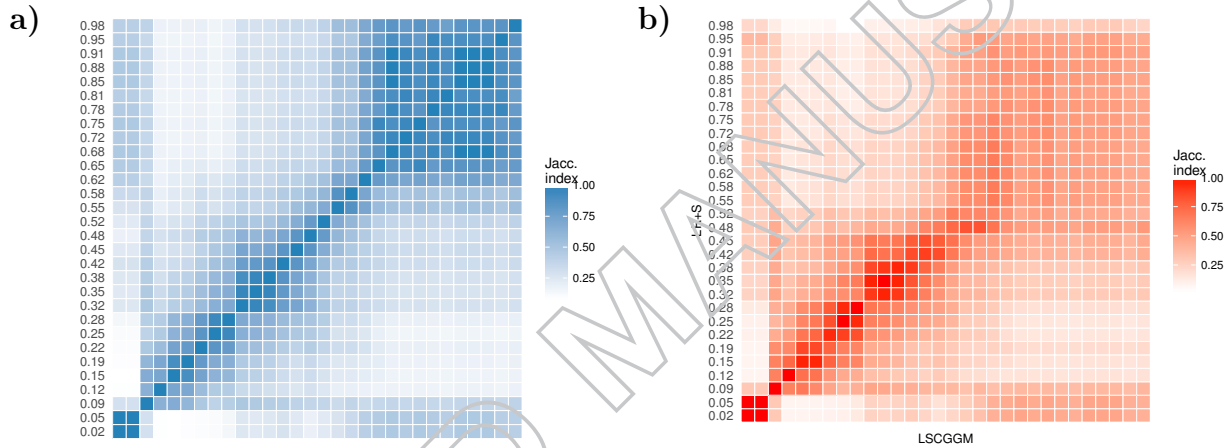

Figure 3: Sensitivity of LSCGGM and LR+S to the tuning parameter γ. For any two graphs, their similarity is computed using the Jaccard Index of their edge sets. (a) Similarities between the edges sets of the graphs returned by LSCGGM in the Child cohort, as a function of γ (for 30 values of $\gamma \in (0.02, 0.98)$). (b) Similarities between the graphs returned by LSCGGM and LR+S in the Child cohort.

28

connections between amino-acids. Here again, a similar result was obtained in the Mother cohort (see suppl. mat.). The supplementary materials also contain the full name of the metabolites being modelled.

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 5a we plot the similarity between graphs estimated at the same value of γ (see suppl. mat. for a plot of

Figure 4: Adjacency matrices of the graphs returned by the LSCGGM and LR+S methods for $\gamma = 0.81$ and $\gamma = 0.68$ respectively.

29

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 Indices 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 $E(V) \leq 1$ (see suppl. mat.).
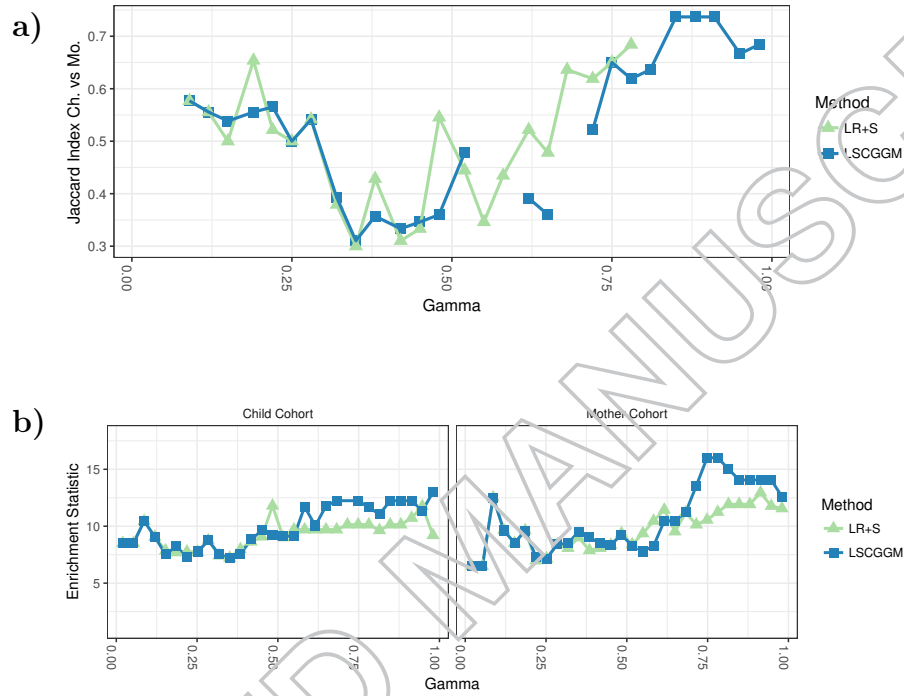

Figure 5: (a) Comparing estimates across cohorts. For each value of γ and each method, we plot the similarity between the estimate obtained in one cohort against the one obtained in the other. We limit ourselves to values of γ for which the estimates in both cohorts comport 15 edges or more. (b) Enrichment statistic, as a function of the tuning parameter γ.

30

Finally, we use the "enrichment statistic" defined earlier. 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 5b shows the value taken by this statistic across cohorts and methods. Associated p-values can be found in the see supplementary materials. 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...

7   Discussion

We discussed the problem of estimating a conditional Gaussian graphical model in the presence of latent variables. Building on the framework introduced by the authors of 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. Among other theoretical results, we established that the proposed approach is well-behaved in the high-dimensional regime. Through simulations and an application to a modern dataset comprising genetic and metabolic measurements, we compared the performances of this approach to alternative methods. In particular, 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.

The rise of high-throughput genetics, along with progress in data linkage, biobanking

31

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 paper, 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.

When it comes to lasso-type estimators, choosing an appropriate value of the tuning parameters can also be challenging. In simulations, our method seems to be 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.

Another limitation of such estimators comes from the fact that consistency/identifiability conditions are highly likely to be violated in real world 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.* instrumental variables). As shown in our application this is an achievable goal.

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

32

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.

Acknowledgements

References

Bach, F. (2008), 'Consistency of trace norm minimization', *J. Mach. Learn. Res.* **8**, 1019–1048.

Bai, Z. & Silverstein, J. (2009), *Spectral Analysis of Large Dimensional Random Matrices*, Springer New York.

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008), 'Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data', *J. Mach. Learn. Res.* **9**, 485–516.

Boyd, A. et al. (2012), 'Cohort profile: The children of the 90s–the index offspring of the avon longitudinal study of parents and children', *International Journal of Epidemiology* **42**(1), 111–127.

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2010), 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Foundations and Trends in Machine Learning* **3**(1), 1–122.

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.

Eckstein, J. & Bertsekas, D. P. (1992), 'On the douglas—rachford splitting method and the proximal point algorithm for maximal monotone operators', *Mathematical Programming* **55**(1-3), 293–318.

Fraser, A. et al. (2012), 'Cohort profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort', *International Journal of Epidemiology* **42**(1), 97–110.

34

Friedman, J., Hastie, T. & Tibshirani, R. (2008), 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics* **9**(3), 432–41.

Goldstein, T. & Osher, S. (2009), 'The split bregman method for l1-regularized problems', *SIAM Journal on Imaging Sciences* **2**(2), 323–343.

Hastings, J. et al. (2012), 'The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013', *Nucleic Acids Research* **41**(D1), D456–D463.

Lauritzen, S. (1996), *Graphical Models*, Clarendon Press.

Li, X., Xie, H., Chen, L., Wang, J. & Deng, X. (2014), 'News impact on stock price return via sentiment analysis', *Knowledge-Based Systems* **69**, 14–23.

Löfberg, J. (2004), Yalmip : A toolbox for modeling and optimization in matlab, *in* 'Proceedings of the CACSD Conference', pp. 284 – 289.

Ma, S., Xue, L. & Zou, H. (2013), 'Alternating direction methods for latent variable gaussian graphical model selection', *Neural Computation* **25**(8), 2172–2198.

Meinshausen, N. & Bühlmann, P. (2010), 'Stability selection', *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **72**(4), 417–473.

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**.

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), 55–80.

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'.

Stearns, F. W. (2010), 'One hundred years of pleiotropy: A retrospective', *Genetics* **186**(3), 767–773.

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.

Vandenberghe, L. & Boyd, S. (1996), 'Semidefinite programming', *SIAM Review* **38**(1), 49–95.

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).

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.', *The annals of applied statistics* **5**(4), 2630–2650.

36

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.

37