Partial correlation based penalty functions and prior distributions for Gaussian graphical models

by

Jack Storror Carter

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

October 2021

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

List of Algorithms

Acknowledgments

First and foremost I would like to thank Jim Smith and David Rossell. Jim's constant enthusiasm meant that I left every meeting with new found energy and optimism. As a mentor he has undoubtedly had a great impact on my early career. I look forward to continuing working with him. David has been like a second supervisor to me and without him this thesis would be nothing like it is now. His advice and guidance have made me a better researcher.

Thank you to my family for always being there if I need them and for getting me to where I am today.

Finally to all the amazing friends I've made during my PhD. Thank you for helping me to enjoy my PhD years and to come out the other side a (hopefully) better and happy person. Special shout outs go to: my colleagues in the department, particularly my fellow OxWaSPers; all my housemates over the years - the Botley crew and the highlanders; the bridge club and all the people I've played sports with, particularly the climbers.

Declarations

I declare that this thesis is my own work that I carried out under the supervision of Professor Jim Smith for the degree of Doctor of Philosophy in Statistics. I have not used sources or means without declaration in the text. I confirm that this thesis has not been submitted for a degree at any other university.

The contents of Chapters 2 and 3 are contained in the article Carter et al. [2021] written during my PhD in collaboration with Dr. David Rossell and Professor Jim Smith which has been submitted and is under review by the Scandinavian Journal of Statistics.

Abstract

Graphical models are a useful tool for encoding conditional independence relations. A common goal is to select the graphical model that best describes the conditional independence relationships between variables given observations of these variables. Under the additional Gaussian assumption, conditional independence is equivalent to zero entries in the inverse covariance matrix Θ. Thus sparse estimation of Θ in turn specifies a graphical model and the associated conditional independencies. Popular frequentist methods for this often involve placing a penalty function on Θ and maximising a penalised likelihood, whilst Bayesian methods require specification of a prior distribution on Θ.

Conditional independence relations are invariant to non-zero scalar multiplication of the variables, however in this thesis we show that essentially all current penalised likelihood methods and many prior distributions are not invariant to such transformations of the variables. In fact many methods are very sensitive to rescaling of the variables which can, and often does, result in a vastly different selected graphical model. To remedy this issue we introduce new classes of penalty functions and prior distributions which are based on partial correlations. We show that such penalty functions and prior distributions lead to scale invariant estimation and posterior inference on Θ.

We pay particular attention to two penalty functions in this class. The partial correlation graphical LASSO places an L_1 penalty on the partial correlations whilst the spike and slab partial correlation graphical LASSO is a penalty function based on a spike and slab prior formulation. The performance of these penalty functions is compared to that of current popular penalty functions in simulated and real world settings. We also investigate spike and slab priors in general for Gaussian graphical models and point out that care must be taken when considering the positive definiteness of Θ. With this in mind we provide some theoretical results based on Wigner matrices.

Chapter 1

Introduction

A graphical model is a statistical model associated to a graph in which the nodes of the graph correspond to random variables of interest. The edges in the graph represent allowed conditional dependencies between the variables, or, more relevantly, the lack of an edge in the graph represents some conditional independence relationship between the associated variables. Informally, conditional independence means that two variables are independent when given the value of some other variable(s). This is a useful property to investigate in many applications, in particular due to the link with causality. This is because different combinations of conditional independence relationships directly determine whether or not some collections of variables can be considered causes of others [Pearl, 2009]. Furthermore, conditional independence between variables rules out a direct causal relationship between them. A graphical model is a tool which is able to encode complex conditional independence relationships between variables and provides a visual representation of these relationships to aid interpretability even when there is a large number of such variables.

The study of graphical models can be split into two categories. First is when the graphical model is given. This might be due to expert or application specific knowledge and potential or known conditional independencies between the variables. Any statistical analysis of the data can then take advantage of these assumed relationships through, for example, the resulting factorisations in the joint density function. The second category is graphical model selection when the graphical model is not given and must be selected using data. This may be done, for example, when one lacks specific knowledge of the relationships between variables and better understanding of potential conditional independencies is desired.

When the variables are assumed to be jointly Gaussian, conditional independence between the variables corresponds exactly to the zero entries in the inverse covariance matrix, also called the precision matrix. In this case graphical model selection is therefore closely related to sparse estimation of the precision matrix. Many methods have been proposed for sparse precision matrix estimation and Gaussian graphical model selection, and perhaps the most well known methods are based on penalised likelihood estimation where a penalty function is added to the log-likelihood. The most prominent penalised likelihood method is the graphical LASSO (GLASSO) which places an L_1 penalty on the entries of the precision matrix.

A key property of conditional independence relationships and therefore graphical models is that they are invariant to rescaling of the variables. That is, if one multiplies the variables element-wise by some vector with non-zero entries, the underlying graphical model remains unchanged as do its associated causal relationships. However, it will be shown in this thesis that essentially all current penalised likelihood methods are not invariant to such rescalings, as well as many Bayesian methods. This issue is well known by many applied researchers who appreciate that when using GLASSO, rescaling of the variables can and often will result in the selection of a vastly different graphical model.

In this thesis we address this issue by proposing a new framework for penalised likelihood and Bayesian methods in Gaussian graphical models which are based on partial correlations. The main contributions of this thesis are:

(i) A novel penalised likelihood framework based on partial correlations that produce estimates and model selection that are invariant to scalar multiplication of the variables. (Chapter 2)

(ii) An investigation of two specific forms of penalty function in this class and their application to both simulated data - when the data generating process is known - and real-world data. (Chapters 2 and 4)

(iii) A novel Bayesian framework for prior distributions based on partial correlations that produce posterior inference that is invariant to rescaling of the variables. (Chapter 3)

(iv) The application of certain appropriate spike and slab prior formulations within this new Bayesian framework. (Chapters 4 and 5)

(v) Some new theory that relates to the positive definiteness of the precision matrix under these spike and slab priors. (Chapter 5)

The content of Chapters 2 and 3 can also be found in Carter et al. [2021], a paper which has been submitted to the Scandinavian Journal of Statistics and is currently under review. Two further papers are planned from the content of this thesis focusing on the performance of the non-convex penalty functions discussed in Chapter 4 and on the theoretical results presented in Chapter 5.

One of the penalised likelihoods we investigate is based on setting an L_1 penalty on the partial correlations and so is directly comparable to the GLASSO. In our simulated results we will show that this new penalised likelihood generally performs better than GLASSO in terms of both estimation and model selection, as well as enjoying the advantage of scale invariance.

In this chapter we begin by introducing conditional independence, graphical models and Gaussian graphical models. In depth discussion of these topics can be found in, for example, Whittaker [1990], Lauritzen [1996] and Studeny [2006]. A more recent and very comprehensive review of graphical models is Maathuis et al. [2018]. The next sections will summarise some key topics in the literature. We will then review current methods for Gaussian graphical model selection before outlining the remainder of the thesis.

1.1 Conditional independence

We begin by formally defining conditional dependence and stating some specific properties of the conditional independence relation. Since graphical models are used to encode conditional independence relationships, it is important that the resulting relationships satisfy these properties. The content of this section and more information can be found in chapter 3 of Lauritzen [1996].

Definition 1. Let X, Y, Z be random variables with a joint distribution $\mathbb{P}$. It is said that X *is conditionally independent of* Y *given* Z *under* $\mathbb{P}$ and it is written $X \perp\!\!\!\perp Y \mid Z\,[\mathbb{P}]$ if for any measurable set A in the sample space of X there exists a version of the conditional probability $\mathbb{P}(A \mid Y, Z)$ which is a function of Z alone.

We assume that the joint distribution $\mathbb{P}$ is fixed and omit this from the notation. When the three variables admit a joint density f with respect to a product measure μ then $X \perp\!\!\!\perp Y \mid Z$ if and only if

$$f_{X,Y|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z)$$

holds almost surely with respect to $\mathbb{P}$. That is, the conditional density of $X, Y \mid Z$ factorises into the two conditional densities of $X \mid Z$ and $Y \mid Z$. In this way we see that conditonal independence is equivalent to independence of the random variables $X \mid Z = z$ and $Y \mid Z = z$ for all possible z.

The conditional independence relation has the following properties, where h denotes an arbitrary measurable function on the sample space of X.

(C1) If $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$.

(C2) If $X \perp\!\!\!\perp Y \mid Z$ and $U = h(X)$, then $U \perp\!\!\!\perp Y \mid Z$.

(C3) If $X \perp\!\!\!\perp Y \mid Z$ and $U = h(X)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$.

(C4) If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (W, Y) \mid Z$.

These properties can be seen as fundamental to the notion of conditional independence as therefore should be adhered to in any conditional independence model.

An additional property of conditional independence is that it is invariant to non-zero scalar multiplication of the variables. Consider vectors a, b, c which are of the same dimension as X, Y, Z respectively and that have non-zero entries. Define the transformed random variables $X' = a^T X$, $Y' = b^T Y$, $Z' = c^T Z$, where M^T denotes the transpose of the matrix (or vector) M, and denote by $\mathbb{P}'$ the joint probability distribution of (X', Y', Z'). Let A' be a measurable set in the sample space of X'. Then $A = \{x : a^T x \in A'\}$ is a measurable set in the sample space of X and $X \in A$ if and only if $X' \in A'$. Hence, by properties of conditional probabilities,

$$\begin{aligned} \mathbb{P}'(A' \mid Y', Z') &= \mathbb{P}(A \mid b^T Y, c^T Z) \\ &= \mathbb{P}(A \mid Y, Z) \end{aligned}$$

It follows that $X \perp\!\!\!\perp Y \mid Z \, [\mathbb{P}]$ if and only if $X' \perp\!\!\!\perp Y' \mid Z' \, [\mathbb{P}']$

1.2 Graphical models

In this section we define a graph $\mathcal{G}$ and explain that a particular relation called *separation* in the graph satisfies analogs of the properties (C1)-(C4). We then go on to demonstrate how a graph can be used to represent conditional independence relationships amongst a group of random variables in a graphical model. The content of this section and more information can be found in chapters 2 and 3 of Lauritzen [1996].

A *graph* is a pair $\mathcal{G} = (V, E)$ where V is called the *vertex set* and E is called the *edge set*. The vertex set V can be any finite set and the elements of V are called *vertices* or *nodes*. The edge set E is a subset of $\{(u, v) : u, v \in V, u \neq v\}$ and the elements of E are called edges.

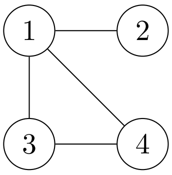


Figure 1.1: A visual representation of the undirected graph $\mathcal{G}$ with vertices $V = \{1, 2, 3, 4\}$ and edges $E = \{(1, 2), (1, 3), (1, 4), (3, 4)\}$

An edge $(u, v) \in E$ is called *undirected* if $(v, u) \in E$ also. If all of the edges in E are undirected then $\mathcal{G}$ is called an *undirected graph.* For an undirected graph the edge set can be simplified by omitting one of $(u, v), (v, u)$. In this thesis we focus on undirected graphs and so for the remainder of this chapter all graphs will be assumed to be undirected.

A graph can be visually represented with the vertices displayed by dots and an edge displayed by a line between the relevant dots. An example of such a visualisation can be seen in Figure 1.1.

Two vertices are called *adjacent* if they have an edge joining them. A *path* of length k from vertex u to vertex v is a sequence $u = u_0, u_1, \ldots, u_k = v$ of distinct vertices such that $(u_{i-1}, u_i) \in E$ or $(u_i, u_{i-1}) \in E$ for all $i = 1, \ldots, k$. A path in which $u = v$ is called a *cycle.* A *decomposable* graph is a graph in which every cycle of length greater or equal to 4 possesses a *chord* - two non-consecutive vertices that are adjacent.

A subset $C \subset V$ is called a (u, v)*-separator* if all paths from u to v intersect C. For subsets $A, B \subset V$, C is said to *separate* A *from* B if C is a (u, v) separator for all $u \in A$ and $v \in B$. For example, in the graph of Figure 1.1 the vertex 1 separates 2 from $\{3, 4\}$.

Define the relation $\overset{\mathcal{G}}{\perp}$ via separation such that $A \overset{\mathcal{G}}{\perp} B \mid C$ if and only if C separates A from B in $\mathcal{G}$. It is fairly straight forward to show that $\overset{\mathcal{G}}{\perp}$ satisfies analogues of the properties (C1)-(C4) where W, X, Y, Z are replaced by subsets of V and the function U is replaced by a subset of its argument. Hence separation in a graph would be a suitable way to represent conditional independence relationships.

In a graphical model, the vertex set V of the graph $\mathcal{G}$ corresponds to the indices of the set of random variables of interest. The edges in the graph represent conditional dependencies between the variables, or more relevantly the lack of an edge represents some conditional independence relationship between the variables. These conditional independencies can be read from the graph via a Markov property - a property that is related to separation.

Let $X = (X^{(1)}, \dots, X^{(p)})$ be a vector of random variables and $\mathcal{G} = (V, E)$ an undirected graph with vertex set $V = \{1, \dots, p\}$. Since $\mathcal{G}$ is undirected and V is a subset of the natural numbers, we only allow E to contain edges of the form (i, j) with $i < j$. The vertex $i \in V$ corresponds to the variable $X^{(i)}$. For a subset $A \subseteq V$ we let $X^{(A)} = (X^{(i)})_{i \in A}$ and $X^{(-A)} = (X^{(i)})_{i \notin A}$. A range of Markov properties have been proposed for encoding conditional independence relationships on X via $\mathcal{G}$. Two common Markov properties are defined as follows.

Definition 2. A probability measure $\mathbb{P}$ for the random variabes X obeys the *pairwise Markov property* relative to $\mathcal{G}$ if for any pair of non-adjacent vertices i, j,

$$X^{(i)} \perp\!\!\!\perp X^{(j)} \mid X^{(-\{i,j\})}.$$

A probability measure $\mathbb{P}$ for the random variabes X obeys the *global Markov property* relative to $\mathcal{G}$ if for any disjoint subsets $A, B, S \subset V$ such that S separates A from B in $\mathcal{G}$,

$$X^{(A)} \perp\!\!\!\perp X^{(B)} \mid X^{(S)}.$$

Generally, the global Markov property is a stronger property than the pairwise Markov property. However, if the distribution of the variables has a positive and continuous density, as is the case for a multivariate Gaussian random vector for example, then the two properties are equivalent [Pearl and Paz, 1987] (also see Theorem 3.7 of Lauritzen [1996]).

Under the graphical model $\mathcal{G}$ the variables X are assumed to satisfy all conditional independencies given by its Markov property. Note that this does not restrict additional conditional independencies not specified by $\mathcal{G}$ from holding. Since the global Markov property corresponds directly to separation in $\mathcal{G}$, the resulting assumed conditional independencies therefore satisfy each of (C1)-(C4).

1.3 Gaussian graphical models

The conditional independence relations given by a graphical model are often a too general framework for data analysis of continuous random variables and so additional assumptions about the joint distribution of the variables are required. In Gaussian graphical models the additional assumption is made that X has a multivariate Gaussian distribution. In this section we will show that a Gaussian graphical model corresponds to zero entries in the precision matrix. We will then discuss how to calculate the maximum likelihood estimate (MLE) of the precision matrix under a specific graphical model. The content of this section and more details can be found in chapter 2 of Whittaker [1990] and chapter 9 of Maathuis et al. [2018].

Let $X = (X^{(1)}, ..., X^{(p)}) \sim \mathrm{N}(\mu, \Sigma)$ be a p-dimensional multivariate Gaussian random vector with unknown mean $\mu \in \mathbb{R}^p$ and $p \times p$, symmetric, positive-definite covariance $\Sigma = (\sigma_{ij})_{i \leq i,j \leq p}$. We denote the precision matrix by $\Theta = (\theta_{ij})_{1 \leq i,j \leq p} = \Sigma^{-1}$ which is also $p \times p$, symmetric and positive definite. For $A, B \subseteq \{1, \ldots, p\}$ we let μ_A, Σ_{AB} and Θ_{AB} denote the corresponding subvector and submatrices. The probability density function of X is written as

$$f_{\mu,\Sigma}(x) = (2\pi)^{-p/2} (\det(\Sigma))^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right),$$

where $\det(A)$ denotes the determinant of the matrix A, or equivalently in terms of the precision matrix as

$$f_{\mu,\Sigma}(x) = (2\pi)^{-p/2} (\det(\Theta))^{1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Theta (x-\mu)\right).$$

A key property of the Gaussian distribution is that it is closed under conditioning (for a proof of this result see Proposition 9.1.1 of Maathuis et al. [2018]).

Proposition 1. *Let A and B partition $\{1, \ldots, p\}$. Then the conditional distribution of $X^{(A)}$ given $X^{(B)}$ is also Gaussian with covariance matrix equal to Θ_{AA}^{-1}.*

This proposition allows a key interpretation of the precision matrix Θ. First consider a singleton $A = \{1\}$ and $B = \{2, \ldots, p\}$. It follows that the conditional distribution of $X^{(1)} \mid X^{(-\{1\})}$ is Gaussian with variance equal to θ_{11}^{-1}. In other words,

$$\theta_{11}^{-1} = \mathrm{Var}\left(X^{(1)} \mid X^{(-1)}\right),$$

and so the diagonal entries of Θ are equal to the inverse partial variances.

Now consider a doubleton $A = \{1, 2\}$ and $B = \{3, \ldots, p\}$. Then the conditional distribution of $X^{(A)} \mid X^{(B)}$ is Gaussian with covariance matrix equal to

$$\Theta_{AA}^{-1} = \frac{1}{\theta_{11}\theta_{22} - \theta_{12}^2} \begin{pmatrix} \theta_{22} & -\theta_{12} \\ -\theta_{12} & \theta_{11} \end{pmatrix}.$$

It follows that the partial covariance between $X^{(1)}$ and $X^{(2)}$ is equal to

$$\mathrm{cov}\left(X^{(1)}, X^{(2)} \mid X^{(-\{1,2\})}\right) = \frac{-\theta_{12}}{\theta_{11}\theta_{22} - \theta_{12}^2},$$

and the partial correlation is equal to

$$\operatorname{corr}\left(X^{(1)}, X^{(2)} \mid X^{(-\{1,2\})}\right) = \frac{-\theta_{12}}{\sqrt{\theta_{11}\theta_{22}}}.$$

Hence, the off-diagonal entries, when rescaled by the relevant diagonal entries, are equal to the negative partial correlations.

Let $\mathcal{G}$ be a graph with vertex set $V = \{1, \ldots, p\}$.

Definition 3. X is said to satisfy the *Gaussian graphical model* relative to $\mathcal{G}$ if $\theta_{ij} = 0$ for all $(i, j) \notin E$

Under the Gaussian graphical model, the graph $\mathcal{G}$ describes the sparsity pattern of the precision matrix Θ, or equivalently specifies certain zero partial correlations between the variables. We will show that X satisfies the Gaussian graphical model relative to $\mathcal{G}$ if and only if it satisfies the global Markov property relative to $\mathcal{G}$.

A well known defining feature of jointly Gaussian random variables is that independence is equivalent to uncorrelatedness. That is, if $X^{(1)}, X^{(2)}$ are jointly Gaussian then $X^{(1)} \perp\!\!\!\perp X^{(2)}$ if and only if the correlation between $X^{(1)}$ and $X^{(2)}$ is zero. This property along with Proposition 1 can be used to show a similar equivalence between conditional independence and zero partial correlations.

Corollary 1. *$X^{(i)} \perp\!\!\!\perp X^{(j)} \mid X^{(-\{i,j\})}$ if and only if $\theta_{ij} = 0$.*

This corollary follows because $X^{(i)} \mid X^{(-\{i,j\})}$ and $X^{(j)} \mid X^{(-\{i,j\})}$ are jointly Gaussian random variables (with distribution only dependent on the conditioned value of $X^{(-\{i,j\})}$ through the mean parameter) with correlation given by $\frac{-\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$.

It then immediately follows that satisfying the Gaussian graphical model relative to $\mathcal{G}$ is equivalent to satisfying the pairwise Markov property relative to $\mathcal{G}$. Since the multivariate Gaussian distribution has a positive and continuous density, this is therefore also equivalent to satisfying the global Markov property.

Now that we have defined the Gaussian graphical model, we turn to the problem of calculating the MLE of Θ under a Gaussian graphical model given observations of X. Suppose we observe n independent samples $(X_1, \ldots, X_n)$ of X and denote their sample covariance by $S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. The pair S and $\bar{X}$ are sufficient statistics for a Gaussian model. The log-likelihood as a function of (μ, Σ) can be written as

$$l(\mu, \Sigma \mid \bar{X}, S) = -\frac{n}{2}\log\left(\det\left(\Sigma\right)\right) - \frac{n}{2}\operatorname{tr}\left(S\Sigma^{-1}\right) - \frac{n}{2}(\bar{X} - \mu)^T\Sigma^{-1}(\bar{X} - \mu) + c,$$

where tr(M) denoted the trace of a matrix M and c is a constant. As a function of (μ, Θ) the log-likelihood is written as

$$l(\mu, \Theta \mid \bar{X}, S) = \frac{n}{2}\log\left(\det\left(\Theta\right)\right) - \frac{n}{2}\mathrm{tr}\left(S\Theta\right) - \frac{n}{2}(\bar{X} - \mu)^T\Theta(\bar{X} - \mu) + c.$$

Assuming that $n > p$, the MLE under the Gaussian graphical model with complete graph $\mathcal{G}$ (i.e. no constraints on Θ) is $\hat{\mu} = \bar{X}$ and $\hat{\Sigma} = S$. For $n \leq p$ the MLE does not exist (i.e. the log-likelihood is unbounded) with probability 1. For a general $\mathcal{G}$ calculation of the MLE is more complicated. First note that the Gaussian graphical model does not put any constraints on μ and so we still have $\hat{\mu} = \bar{X}$. Denoting the set of symmetric, positive definite $p \times p$ matrices by $\mathcal{S}$, the MLE problem then reduces to

$$\begin{aligned} \max_{\Sigma \in \mathcal{S}} \quad & -\log\left(\det\left(\Sigma\right)\right) - \mathrm{tr}\left(S\Sigma^{-1}\right) \\ \text{subject to} \quad & \left(\Sigma^{-1}\right)_{ij} = 0 \text{ for all } (i,j) \notin E \end{aligned}$$

or in terms of Θ to

$$\begin{aligned} \max_{\Theta \in \mathcal{S}} \quad & \log\left(\det\left(\Theta\right)\right) - \mathrm{tr}\left(S\Theta\right) \\ \text{subject to} \quad & \theta_{ij} = 0 \text{ for all } (i,j) \notin E \end{aligned} \tag{1.1}$$

When considered in terms of Θ, this can be shown to be a convex optimisation problem. The dual problem to (1.1) can be shown to be

$$\begin{aligned} \min_{\Sigma \in \mathcal{S}} \quad & -\log\left(\det\left(\Sigma\right)\right) - p \\ \text{subject to} \quad & \Sigma_{ij} = S_{ij} \text{ for all } i = j \text{ or } (i,j) \in E \end{aligned} \tag{1.2}$$

The MLE is not guaranteed to exist for all S and all $\mathcal{G}$, namely because the objective function may be unbounded on the feasible region. A sufficient condition for the MLE to exist is $n > p$. Without placing additional assumptions on the form of $\mathcal{G}$, one cannot obtain a stronger sufficient condition - for example when $\mathcal{G}$ is the complete graph, the MLE exists if and only if $n > p$. However, for certain graphs $\mathcal{G}$, the MLE may exist for smaller n. For more details of this see Section 9.5 of Maathuis et al. [2018].

Assuming the MLE does exist, since this is a convex optimisation problem it can be solved in polynomial time by an interior point method [Boyd and Vandenberghe, 2004]. An even simpler approach, and one that is generally effective, is using a coordinate descent algorithm. To do this, begin with $\Sigma^0 = S$ and iteratively update each of the entries Σ_{ij}, $i \neq j$, $(i,j) \notin E$ by maximising the log-likelihood with all other entries fixed. Note that only the indices $(i,j) \notin E$ need be considered because the other entries for the MLE of Σ are given in (1.2). A similar coordinate descent algorithm can be made for Θ beginning at $\Theta^0 = I$ the identity matrix.

1.4 Gaussian graphical model selection

Up to this point we have assumed that the graph $\mathcal{G}$ specifying the graphical model is given. This might be the case when certain prior or expert knowledge informs conditional independencies that are known to exist between the variables. However, in many applications this will not be the case and instead one may wish to select a graphical model given data. Under the assumption that X follows a Normal distribution with precision matrix Θ, the goal is therefore to identify the graph $\mathcal{G}$ with edge set given by $(i,j) \in E \iff \theta_{ij} = 0$, based on data summarised by the sample size n and sample covariance S. In this section we introduce some methods for doing this.

A naive approach outlined by Whittaker [1990] is to initially estimate Θ by S^{-1} (provided the inverse exists which occurs with probability 1 when $n > p$). From S^{-1} the sample partial correlations can be obtained by scaling the matrix to have unit diagonals. One may then choose a threshold c for which any sample partial correlations in $(-c, c)$ are set to be equal to 0. The graphical model is then selected based on these zero entries.

Note that this procedure bases edge inclusion on the absolute value of the sample partial correlations, rather than the off-diagonal entries of S^{-1}. This will be a key theme through the remainder of the thesis. However, this approach has a number of problems. First, there is no obvious and unequivocal way to set the threshold c. Second, the absolute value of the sample partial correlations does not necessarily correspond to the likelihood of two variables being conditionally dependent. Third, this ignores any dependence between the sample partial correlations - fixing one partial correlation to be zero may change the MLE of another partial correlation. Finally, the resulting estimate is not guaranteed to be positive definite.

There are various expedient tools that can mitigate some of these problems, however these are ad hoc and any one difficult to justify over another. For example, we could consider the following stepwise backward-search algorithm (or similarly a forward-search algorithm) which is similar to that suggested in Højsgaard et al. [2012]. Let $\mathcal{G}_0$ be the complete graph and Θ_0 the MLE under the Gaussian graphical model $\mathcal{G}_0$. Then define the graph $\mathcal{G}_1$ which is equal to $\mathcal{G}_0$ except with the edge associated to the smallest estimated partial correlation in absolute value in Θ_0 removed. Then let Θ_1 be the MLE under $\mathcal{G}_1$. This procedure can be continued until either a desired level of sparsity is reached, or until the empty graph is reached. A single graphical model from this sequence can then be selected by using some selection criterion, for example the Bayesian information criterion. However, this procedure can become computationally expensive for large p.

This stepwise procedure is similar to another class of methods called model search algorithms. Model search algorithms compare some subset of all possible graphical models and select between them using some criterion. A model search algorithm is defined by its method for selecting this subset of models. Ideally one may consider the set of all possible graphical models. However, this is generally computationally infeasible for even moderate problem size. If there are p variables then there are $\frac{1}{2}p(p-1)$ possible edges in a graphical model and $2^{\frac{1}{2}p(p-1)}$ possible graphical models. Even for $p = 5$ this gives 1024 models to consider. Hence for even moderately sized problems, any model search algorithm must not consider all models, and therefore may potentially miss models that describe the data well.

1.4.1 Penalised likelihood

Many approaches to Gaussian graphical model selection instead focus on sparse estimation of Θ and then select the model that matches this sparse estimate. One popular frequentist method for sparse estimation is the maximisation of a penalised likelihood. Quite simply, this adds a penalty term to the log-likelihood function which penalises non-zero parameters and therefore encourages sparsity in the estimate. Penalised likelihood approaches are common in linear regression, the most famous being the LASSO of Tibshirani [1996], and many of these approaches have been adapted for application to Gaussian graphical models. These adaptations come in two major forms.

The first form is when the penalty is applied directly to the precision matrix Θ. Estimation of Θ then simply involves maximisation of a penalised likelihood of the form

$$l(\Theta \mid S) - Pen(\Theta)$$

where

$$l(\Theta \mid S) = \frac{n}{2}\left(\log\left(\det\left(\Theta\right)\right) - \operatorname{tr}\left(S\Theta\right)\right)$$

is the log-likelihood function for Θ after removing constants and $Pen(\Theta)$ is a penalty function.

Penalty functions are often chosen to be increasing in $|\theta_{ij}|$ with a local minimum at $\theta_{ij} = 0$. In this way, smaller estimates of the θ_{ij} are encouraged - this is commonly referred to as regularisation of the θ_{ij}. Choosing a penalty function which is non-differentiable at $\theta_{ij} = 0$ also allows the possibility of exact zero estimates.

The LASSO penalty assigns an L_1 penalty to the coefficients in a linear regression. This has been adapted to the Gaussian graphical model setting in the graphical LASSO (GLASSO) proposed and investigated by Yuan and Lin [2007], Friedman et al. [2008] and Banerjee et al. [2008]. Under the GLASSO the penalty function is of the form

$$Pen(\Theta) = \rho \sum_{i,j} |\theta_{ij}|$$

where $\rho > 0$ is called the penalty or regularisation parameter. The objective function for the GLASSO is concave and therefore benefits from the advantages of convex optimisation to allow estimates to be obtained very quickly and efficiently. Furthermore, unlike the regular MLE for Θ which only exists when $n > p$, the penalised likelihood estimate of GLASSO exists for any sample size n. It was also shown by Yuan and Lin [2007] that the GLASSO estimate is equivalent to maximising the log-likelihood under the constraint that the L_1 norm of Θ is bounded by a certain amount.

It was noted in linear regression that the LASSO tends to induce a significant bias on large, non-zero regression coefficients. This is due to the penalty increasing linearly in $|\theta_{ij}|$ and therefore inflicting a large penalty on large $|\theta_{ij}|$ and, more importantly, the gradient of the penalty being constant in $|\theta_{ij}|$. To reduce this bias, non-convex penalty functions have been proposed. Two examples of these are the smoothly clipped absolute deviation (SCAD) penalty, proposed by Fan and Li [2001] and adapted to Gaussian graphical models by Fan et al. [2009] and the minimax concave penalty (MCP) proposed by Zhang [2010].

The SCAD penalty is of the form

$$\sum_{i,j} \mathrm{SCAD}_{\lambda,a}(|\theta_{ij}|)$$

where

$$\mathrm{SCAD}'_{\lambda,a}(x) = \lambda \left(\mathbb{I}(x \leq \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} \mathbb{I}(x > \lambda) \right)$$

for $x > 0$. Here $\mathbb{I}$ denotes the indicator function and $(x)_+$ denotes the maximum of x and 0.

The MCP is of a similar form to the SCAD penalty replacing $\mathrm{SCAD}_{\lambda,a}$ with

$\mathrm{MCP}_{\lambda,a}$ where

$$\mathrm{MCP}'_{\lambda,a}(x) = \left(\lambda - \frac{x}{a}\right) \mathbb{I}(x \leq a\lambda).$$

The SCAD penalty has constant gradient equal to λ around 0 before decreasing linearly to 0. The derivative of the MCP instead decreases linearly to 0 straight away. For both the SCAD penalty and the MCP we refer to $\lambda > 0$ as the regularisation parameter.

A second form of Gaussian graphical model selection using penalised likelihoods occurs due to a direct connection to linear regression. By regressing a single variable $X^{(i)}$ on the remaining variables $X^{(-i)}$, the resulting regression coefficients correspond to the entries in the ith row of Θ through $\beta_j = -\sigma_{ii}\theta_{ij}$. This relationship motivated the method of Meinshausen and Bühlmann [2006] which implemented successive linear regressions on each of the coefficients using the LASSO. The zero regression coefficients then informed the selected graphical model. For additional information see, for example, Section 12.3.3 of Maathuis et al. [2018].

Theoretical results for model selection and so called 'oracle' properties exist for both GLASSO and the method of Meinshausen and Bühlmann [2006] under certain conditions on the true underlying Θ, the sample size n and the problem size p. If these conditions are satisfied, both methods are proven to return the true underlying graphical model with a certain probability, for certain choices of the regularisation parameter. Oracle properties relate to the distance between the true and estimated Θ being bounded. Again, under certain conditions the GLASSO and Meinshausen estimates satisfy oracle properties on the l_1-norm, for certain choices of the regularisation parameter. For more information on these properties see Section 12.3 and Section 14.1 of Maathuis et al. [2018].

Parameter selection is an important aspect of penalised likelihood methods. For the SCAD penalty and the MCP default values of $a = 3.7$ and $a = 2$ have been proposed, respectively. For the regularisation parameter is it common to calculate penalised likelihood estimates for a sequence of parameters and then choose between them via some selection criterion. Popular choices of selection criterion are cross validation, the Bayesian information criterion (BIC) and the extended Bayesian information criterion (EBIC). Additional details on the BIC will be given in Section 2.2, an overview of the BIC and cross validation can be found in Lian [2011] and details on the EBIC in Foygel and Drton [2010]. The SCAD penalty has been shown to achieve consistent model selection when parameter selection is via the BIC [Lian, 2011; Gao et al., 2012]. When instead predictive power is desired over model selection, cross validation offers good performance - see Vujačić et al. [2015].

1.4.2 Bayesian methods

A standard Bayesian procedure for model selection requires specification of prior distributions on Θ given a graphical model, $\pi_1(\Theta \mid \mathcal{G})$, and on the model space, $\pi_2(\mathcal{G})$. The most common choice for model space prior is a discrete uniform prior, assigning each possible graph equal prior probability. However, such a prior heavily favours graphs of moderate size where the size of the graph refers to the number of edges. For example, when $p = 5$ there are 1024 possible graphical models, but only one graph with no edges and one complete graph with 10 edges. Hence each of these have a $\frac{1}{1024}$ probability under the uniform prior. On the other hand, there are 252 graphs with 5 edges and so under the uniform prior this has probability of approximately $\frac{1}{4}$.

One alternative is to set

$$\pi_2(\mathcal{G}) = \eta^{\text{size}(\mathcal{G})}(1-\eta)^{m-\text{size}(\mathcal{G})}$$

where m is the maximum size of $\mathcal{G}$ and $\eta \in (0,1)$. This assumes that the inclusion probability of any edge is constant and equal to η and allows for favouring of more sparse graphs. Another option is to separately set a prior on the model size and on the model given the model size. This allows for even more flexible prior specification.

The prior for the joint space $(\Theta, \mathcal{G})$ is given by $\pi(\Theta, \mathcal{G}) = \pi_1(\Theta \mid \mathcal{G})\pi_2(\mathcal{G})$. Given samples of X, which are summarised by the sample size n and the sample covariance S, the resulting posterior density is equal to

$$\pi(\Theta, \mathcal{G} \mid n, S) \propto L(\Theta \mid S, n)\pi_1(\Theta \mid \mathcal{G})\pi_2(\mathcal{G})$$

where the likelihood function L depends on $(\Theta, \mathcal{G})$ only through Θ. For model selection, we are interested in the posterior density of $\mathcal{G}$, which requires integrating the full posterior with respect to Θ over the space of symmetric, positive definite matrices. Thus we need to calculate

$$\begin{aligned}\pi(\mathcal{G} \mid n, S) &= \int_{\mathcal{S}} \pi(\Theta, \mathcal{G} \mid n, S)\, d\Theta \\ &\propto \pi_2(\mathcal{G}) \int_{\mathcal{S}} L(\Theta \mid S, n)\pi_1(\Theta \mid \mathcal{G})\, d\Theta \qquad (1.3)\end{aligned}$$

In an ideal world, this posterior probability would be calculated for all possible $\mathcal{G}$. If desired, one may then select a number of high posterior probability models to give an idea of the conditional independencies that the data suggests. However, this approach has two problems. First, calculation of these probabilities is not necessarily straight forward due to the integral. Second, as with the frequentist case, calculation of the posterior probability for all $\mathcal{G}$ is often infeasible, even for moderate p. Hence, some search algorithm is needed to traverse the model space and identify graphs that are likely to have high posterior probability. One common tool for comparing two models is the Bayes factor which gives the ratio of posterior probabilities for two models.

One approach to this first issue is to choose prior densities $\pi_1(\Theta \mid \mathcal{G})$ which allow easy calculation of this integral. This may be done through the graphical Wishart (G-Wishart) prior on Θ, or equivalently a hyper-inverse Wishart prior on Σ. Introduced by Dawid and Lauritzen [1993], the hyper-inverse Wishart distribution is conjugate for Σ under a Gaussian graphical model and satisfies the Gaussian graphical model with probability 1. Although the hyper-inverse Wishart is only defined for decomposable graphs, the G-Wishart can be generalised to non-decomposable graphs. As long as the graph $\mathcal{G}$ is decomposable, the G-Wishart prior allows closed form calculation of the integral in (1.3), and therefore the posterior probabilities can be calculated up to a normalising constant. Hence Bayes factors can be calculated in closed form since the normalising constant cancels out. These Bayes factors can then be used to explore the space of graphical models by using, for example, a Metropolis-Hastings algorithm as in Madigan et al. [1995] or a reversible jump algorithm as in Giudici and Green [1999] and Dobra et al. [2011].

One disadvantage of these types of algorithms is that they take a long time to explore the whole model space since they only add or remove a single edge in each iteration. The birth and death MCMC algorithm of Mohammadi and Wit [2015] improved on this by allowing more general jumps through the model space. An even greater improvement would be to remove the need for traversing the model space by sampling directly from the posterior of Θ (which then implies the graphical model through its sparsity pattern). Obtaining such samples, or constructing an MCMC algorithm without conditioning on the model $\mathcal{G}$ is challenging, however, due to the distribution being non-continuous with point masses at 0. Instead one may consider a continuous relaxation of the prior on Θ to allow for easier posterior sampling. Some methods for doing this will be introduced later in the thesis.

For more information on Bayesian methods for Gaussian graphical model selection see Chapter 10 of Maathuis et al. [2018].

1.4.3 Other methods

Although in this thesis we focus on penalised likelihood and Bayesian methods for Gaussian graphical model selection, many other methods have been proposed. One of the most prominent alternative methods is the constrained l_1 minimisation for inverse matrix estimation (CLIME) method of Cai et al. [2011]. The CLIME method solves an alternative constrained optimisation problem which instead minimises the L_1 norm of Θ under the constraint that the largest entry of $S\Theta - I$ is bounded at a chosen threshold. This optimisation problem is convex so allows efficient computation. However, CLIME does not guarantee sparsity in the estimate of Θ and so graphical model selection is conducted by thresholding the resulting estimate.

The sparse partial correlation estimation (SPACE) method of Peng et al. [2009] expands on the nodewise regression method of Meinshausen and Bühlmann [2006]. They instead focus on estimation of the partial correlations, rather than the entries of Θ, and conduct the regressions concurrently and dependently such that the resulting regression coefficients avoid certain logical fallacies that may occur in the method of Meinshausen and Bühlmann [2006]. The SPACE method was shown to work particularly well in the presence of hub variables - variables associated to many edges in the graphical model.

Methods based on the score matching loss have been proposed by Forbes and Lauritzen [2015] and Lin et al. [2016]. Instead of focusing on the log-likelihood function, score matching methods instead aim to minimise some score function which quantifies the accuracy of a predictive distribution given a realised value. These are structurally similar to the SPACE method, benefit from convenient computation, are robust to non-Gaussian data and tend to perform well in high dimensional settings.

1.5 Thesis outline

The remainder of the thesis is structured as follows. In Chapter 2 we introduce a class of penalty functions, which the GLASSO penalty, SCAD penalty and MCP all belong to, and point out a fundamental flaw with penalised likelihood estimates based on such penalty functions - namely that scalar multiplication of the variables results in different estimates of Θ and different graphical model selection. We also introduce a new class of penalty functions based on partial correlations and show that these benefit from estimates that are invariant to scalar multiplication. Particular attention is paid to one novel penalised likelihood method - the partial correlation GLASSO (PC-GLASSO) - which places an L_1 penalty on the partial correlations, and we propose a coordinate descent algorithm for calculation of the PC-GLASSO estimates. This chapter concludes with applications on both simulated and real data sets comparing the PC-GLASSO to the GLASSO. The real data sets investigated involve gene expression measurements of colon cancer patients and S&P 500 index stock prices.

Chapter 3 introduces a similar framework based on partial correlations, but this time for prior distributions. We prove a stronger result that the whole posterior distribution is invariant to scalar multiplication of the variables under such priors. These prior distributions are related to penalty functions and we introduce the prior related to the PC-GLASSO. We then compare this PC-GLASSO prior to the GLASSO prior of Wang [2012].

In Chapter 4 we use a spike and slab prior framework to inspire a new penalty function on the partial correlations - the spike and slab PC-GLASSO (SS-PC-GLASSO). This induces a non-convex penalty which, similarly to the SCAD penalty and MCP, aims to reduce the bias on large partial correlations associated to the L_1 penalty of the PC-GLASSO. We compare this penalty to the SCAD penalty and the MCP before proposing methods for parameter selection and computation. The SS-PC-GLASSO is then compared to PC-GLASSO, SCAD and MCP in simulated and real world examples.

Chapter 5 more extensively explores the use of spike and slab priors for Gaussian graphical models. A key observation is made related to the positive definiteness of Θ and the interpretability of the spike and slab prior. We present a theorem, whose proof is based on the theory of Wigner matrices, to combat this issue and provide a strategy for setting parameter values to ensure an interpretable prior. We then discuss choices for spike and slab densities and strategies for posterior inference.

We conclude the thesis in Chapter 6 with a discussion and some key points for major future projects based on the work of this thesis.

Chapter 2

Partial correlation graphical LASSO

We begin this chapter with a motivating example demonstrating a potentially troublesome feature in the GLASSO and other common penalty functions - a feature which we will modify with a newly proposed class of penalty functions containing our novel partial correlation graphical LASSO (PC-GLASSO) method for estimating precision matrices for Gaussian graphical models.

Example The goal of this example is to estimate the precision matrix Θ associated to a p-variate Gaussian random vector. We set $p = 50$ and generate $n = 100$ independent Gaussian draws with zero mean and covariance $\Sigma = \Theta^{-1}$, where Θ follows the so-called star pattern, with $\theta_{ii} = 1$ and $\theta_{i1} = \theta_{1i} = -1/\sqrt{p}$ for $i = 2, \ldots, p$, and $\theta_{ij} = 0$ otherwise.

$$\Theta = \begin{pmatrix} 1 & -\frac{1}{5\sqrt{2}} & -\frac{1}{5\sqrt{2}} & \cdots & -\frac{1}{5\sqrt{2}} \\ -\frac{1}{5\sqrt{2}} & 1 & 0 & \ldots & 0 \\ -\frac{1}{5\sqrt{2}} & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{5\sqrt{2}} & 0 & 0 & \ldots & 1 \end{pmatrix}; \quad \Sigma = \begin{pmatrix} 50 & 5\sqrt{2} & 5\sqrt{2} & \ldots & 5\sqrt{2} \\ 5\sqrt{2} & 2 & 1 & \ldots & 1 \\ 5\sqrt{2} & 1 & 2 & \ldots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 5\sqrt{2} & 1 & 1 & \ldots & 2 \end{pmatrix}.$$

In this setup recovery of the graphical model is relatively straightforward using GLASSO, see for example Yuan and Lin [2007]. Indeed, the top left panel in Figure 2.1 shows the regularisation path for the estimated partial correlations under GLASSO. For a large range of values for the regularisation parameter ρ the truly zero θ_{ij}'s are completely separated from the non-zeroes.

However, suppose that a data scientist decides to first standardise the data to have unit sample variances before applying GLASSO, as is common practise and has been recommended by, for example, Yuan and Lin [2007]. To standardise the data one can replace the sample covariance matrix S by the sample correlation matrix

$$R = \mathrm{diag}(S)^{-1/2} S \mathrm{diag}(S)^{-1/2}$$

where $\mathrm{diag}(S)$ is the diagonal $p \times p$ matrix with diagonal entries equal to those of S. The top right panel of Figure 2.1 shows the regularisation path for the estimated partial correlations when GLASSO is applied to the standardised data. It is clear here that the inference has suffered, and in particular the true graphical model is not recovered for any ρ.

Although not equivalent, it is useful to consider the standardised data as being similar to a Gaussian sample with covariance matrix $\tilde{\Sigma} = \tilde{\Theta}^{-1}$ given by

$$\tilde{\Theta} = \begin{pmatrix} 50 & -\sqrt{2} & -\sqrt{2} & \dots & -\sqrt{2} \\ -\sqrt{2} & 2 & 0 & \dots & 0 \\ -\sqrt{2} & 0 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\sqrt{2} & 0 & 0 & \dots & 2 \end{pmatrix}; \quad \tilde{\Sigma} = \begin{pmatrix} 1 & \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} & \dots & \frac{1}{2}\sqrt{2} \\ \frac{1}{2}\sqrt{2} & 1 & \frac{1}{2} & \dots & \frac{1}{2} \\ \frac{1}{2}\sqrt{2} & \frac{1}{2} & 1 & \dots & \frac{1}{2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\sqrt{2} & \frac{1}{2} & \frac{1}{2} & \dots & 1 \end{pmatrix}$$

We highlight two key differences between Θ, Σ and $\tilde{\Theta}, \tilde{\Sigma}$ which might explain the changes in performance of GLASSO. First, the diagonal entries of $\tilde{\Theta}$ are not all equal, unlike the diagonal entries of Θ. Second, the entries of Σ related to an edge are equal to $5\sqrt{2}$ and the entries not related to an edge are equal to 1. Meanwhile in $\tilde{\Sigma}$ the entries related to an edge are equal to $\frac{1}{2}\sqrt{2}$ and those not related to an edge are equal to $\frac{1}{2}$. The entries related to an edge are much larger, in both absolute and relative terms, in Σ than in $\tilde{\Sigma}$. We conjecture that a combination of these two factors leads to the decreased performance, something that will be explained further later in the chapter.

As a further example, we now multiply the samples related to the second variable by 10 so that the sample is now from a Gaussian distribution with covariance matrix $\bar{\Sigma} = \bar{\Theta}^{-1}$ given by

$$\bar{\Theta} = \begin{pmatrix} 1 & -\frac{1}{50\sqrt{2}} & -\frac{1}{5\sqrt{2}} & \dots & -\frac{1}{5\sqrt{2}} \\ -\frac{1}{50\sqrt{2}} & \frac{1}{100} & 0 & \dots & 0 \\ -\frac{1}{5\sqrt{2}} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{5\sqrt{2}} & 0 & 0 & \dots & 1 \end{pmatrix}; \quad \bar{\Sigma} = \begin{pmatrix} 50 & 50\sqrt{2} & 5\sqrt{2} & \dots & 5\sqrt{2} \\ 50\sqrt{2} & 200 & 10 & \dots & 10 \\ 5\sqrt{2} & 10 & 2 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 5\sqrt{2} & 10 & 1 & \dots & 2 \end{pmatrix}.$$

Now the second diagonal entry of $\bar{\Theta}$ is much smaller than the others, and the entries in the second row and column of $\bar{\Sigma}$ are inflated. In this case, as can be seen in the bottom left panel of Figure 2.1, for many values of ρ the GLASSO estimate only includes edges related to the second variable, even though this is far from the true model. This suggests that GLASSO favours edges related to small diagonal entries in Θ and large off-diagonal entries in Σ.

These examples highlight two key issues with GLASSO. First, the estimates obtained by GLASSO depend on the scale on which the variables are measured. In other words, the estimate and model selected by GLASSO is not invariant to scalar multiplication of the variables. This issue is not restricted to GLASSO but, as will be shown later in the chapter, affects essentially all common penalty functions. Second, under certain scalings of the variables GLASSO can provide inferior and arguably illogical estimates of the precision matrix. As was highlighted in the example, this can occur even when the data has been standardised.

In order to combat these issues, we propose a new class of penalty functions based on partial correlations and investigate one specific penalty function in this class which we call partial correlation graphical LASSO (PCGLASSO). The regularisation path for the above example can be found in Figure 2.1 along with the Kullback-Leibler loss associated to estimates along the regularisation path for PC-GLASSO, GLASSO, SCAD and MCP under data standardised by S. This demonstrates PC-GLASSO's improved performance in estimation of Θ.

The rest of the chapter is organised as follows. Section 2.1 sets notation and reviews popular classes of likelihood penalties which we refer to as *regular* penalty functions. Section 2.2 introduces a new class of penalties on partial correlations, and the PC-GLASSO as a particular case. Section 2.3 briefly specifies two forms of standardising Gaussian data and Section 2.4 compares the GLASSO and PC-GLASSO estimates in the $p = 2$ case. Section 2.5 shows that the PC-GLASSO, as well as the logarithmic and L_0 penalties are scale invariant, while regular penalty functions are not. Section 2.6 informally discusses potential reasons for the poor performance of GLASSO seen in the above example under certain scalings, while Section 2.7 attempts to formalise these ideas with a notion of exchangeable inference. Section 2.8 gives a brief discussion on the penalisation of the diagonal entries of Θ and Section 2.9 discusses computational issues for the PC-GLASSO and gives a certain conditional convexity result. Section 2.10 shows examples on simulated, gene expression and stock market datasets. We end the chapter with a short discussion.

We also note that content from this chapter and the subsequent chapter appear in a paper written with David Rossell and available on arXiv [Carter et al.,

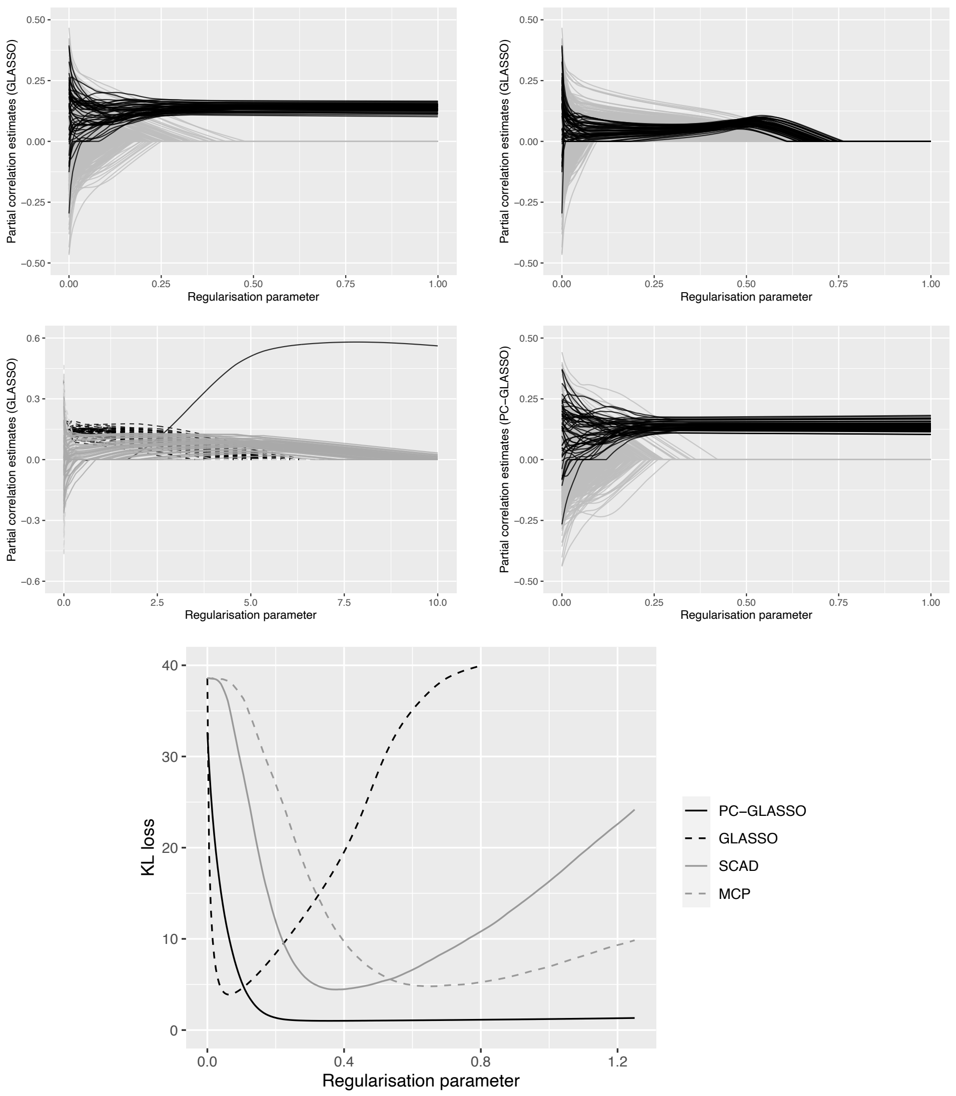


Figure 2.1: Top: Partial correlation regularisation paths for GLASSO in the $p = 50$ star graph example on the original data (left), and standardised data (right). Estimates of truly non-zero θ_{ij} are in black. Middle: Partial correlation regularisation paths for GLASSO (left) when second variable has been multiplied by 10. Partial correlations not related to the second variable are dashed. Partial correlation path for PC-GLASSO (right). Bottom: KL loss over the regularisation paths for different penalties applied to standardised data.

2021] (submitted to Scandinavian Journal of Statistics). The PC-GLASSO method and all results in Sections 2.5 and 2.7 are, to the best of our knowledge, novel.

2.1 Penalised likelihood in Gaussian graphical models

Let $X = (X^{(1)}, ..., X^{(p)}) \sim \mathrm{N}(\mu, \Sigma)$ be a p-dimensional multivariate Gaussian random vector with unknown mean $\mu \in \mathbb{R}^p$ and $p \times p$ positive-definite covariance $\Sigma = (\sigma_{ij})_{i \leq i,j \leq p}$. Suppose we observe n independent samples $(X_1, \ldots, X_n)$ of X and denote their sample covariance by $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^{\mathrm{T}}$, where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. Our goal is to estimate the precision matrix $\Theta = (\theta_{ij})_{1 \leq i,j \leq p} = \Sigma^{-1}$.

A common assumption in Gaussian graphical models is that the data generating process is governed by a sparse undirected graph so that Θ is a sparse matrix with many zero entries, and we have a particular interest in the location of its zero entries. This is due to the equivalence between zero partial covariances and conditional independencies in Gaussian graphical models. The most common frequentist approach to sparse estimation is to maximise a penalised likelihood function of the form $l(\Theta \mid S) - Pen(\Theta)$, where

$$l(\Theta \mid S) = \frac{n}{2}\left[\log(\det(\Theta)) - \mathrm{tr}(S\Theta) - p\log(2\pi)\right], \tag{2.1}$$

is the log-likelihood function, $Pen(\Theta)$ some penalty function and $\mathrm{tr}(A)$ the trace of A. Most popular choices (discussed below) consider penalties that are additive and monotone in $|\theta_{ij}|$, which we refer to as *separable penalties*, and in particular the subclass of penalties differentiable everywhere other than zero, which we refer to as *regular penalties.*

Definition 4. A penalty function $Pen(\Theta)$ is *separable* if

$$Pen(\Theta) = \sum_{i \leq j} pen_{ij}(\theta_{ij}),$$

where $pen_{ii} : (0, \infty) \to \mathbb{R}$ and $pen_{ij} : \mathbb{R} \to \mathbb{R}$ are non-decreasing in θ_{ii} and $|\theta_{ij}|$ respectively for all i and $i < j$.

A separable penalty is *regular* if $pen_{ii} = pen_{jj}$ for all (i, j) and, for all $i < j$, pen_{ij} does not depend on (i, j), is symmetric about 0 and differentiable away from 0.

Most popular penalty functions used for Gaussian graphical models are regular. The GLASSO is a prominent example using an L_1 penalty to produce the point estimate

$$\Theta^{\rho}_{\text{GLASSO}}(S) = \arg\max \log(\det(\Theta)) - \text{tr}(S\Theta) - \rho \sum_{i=1}^{p} \sum_{j=1}^{p} |\theta_{ij}| \tag{2.2}$$

for some given regularization parameter $\rho \geq 0$. This corresponds to the regular penalty function with diagonal penalty $pen_{ii}(\theta_{ii}) = \frac{n}{2}\rho\theta_{ii}$ and off-diagonal penalty $pen_{ij}(\theta_{ij}) = n\rho|\theta_{ij}|$. Note that the penalty function here is multiplied by the sample size n; this is so that the resulting maximisation problem doesn't depend on n and so that similar ranges of values for the parameter ρ are sensible regardless of n. See Meinshausen and Bühlmann [2006] for an alternative that places L_1 penalties on the full conditional regression of each $X^{(i)}$ given $X^{-(i)}$, Banerjee et al. [2008] for computational methods based on parameterising (2.2) in terms of Σ and Yuan and Lin [2007] for a variation that omits the diagonal of Θ from the penalty. Other popular regular penalties include the SCAD penalty [Fan and Li, 2001; Fan et al., 2009] and the MCP penalty [Zhang, 2010; Wang et al., 2016], which were proposed to reduce bias in the estimation of large entries in Θ relative to the L_1 penalty.

Another notable regular penalty is the L_0 penalty

$$Pen(\Theta) = \rho \sum_{i<j} \mathbb{I}(\theta_{ij} \neq 0), \tag{2.3}$$

where $\mathbb{I}$ is the indicator function.

The adaptive LASSO [Zhou et al., 2009; Fan et al., 2009] is an important example of a non-regular penalty. It uses a weighted L_1 penalty where weights depend on the data via some initial estimate of Θ, and hence does not satisfy Definition 4. However, as noted by Bühlmann and Meier [2008] and Candès et al. [2008], the adaptive LASSO can be seen as a first-order approximation of the logarithmic penalty where $pen_{ij}(\theta_{ij}) = \rho \log(|\theta_{ij}|)$, which is regular. Both papers propose an iterative version of adaptive LASSO that formally targets this logarithmic penalty.

2.2 Partial Correlation Graphical LASSO

We propose basing penalties on a reparameterisation of Θ in terms of the (negative) partial correlations

$$\Delta_{ij} := \frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}} = -\text{corr}\left(X^{(i)}, X^{(j)} \mid X^{-(ij)}\right).$$

where $X^{-(ij)}$ denotes the vector X after removing $X^{(i)}$ and $X^{(j)}$.

The precision matrix can be decomposed as $\Theta = \theta^{\frac{1}{2}}\Delta\theta^{\frac{1}{2}}$, where $\theta = \text{diag}(\Theta)$ and Δ is the matrix with unit diagonal and off-diagonal entries Δ_{ij}. The penalised likelihood function then becomes

$$\frac{n}{2}\left[\log(\det(\Delta)) + \sum_i \log(\theta_{ii}) - \text{tr}(S\theta^{\frac{1}{2}}\Delta\theta^{\frac{1}{2}})\right] - Pen(\theta, \Delta). \tag{2.4}$$

We believe that partial correlations are a better measure of dependence than the off-diagonals θ_{ij}, in that they are easier to interpret and invariant to scalar multiplication of the variables. We now introduce a class of additive penalties in this parameterisation, a corresponding prior class, and subsequently state our PC-GLASSO as a particular case.

Definition 5. A penalty *Pen* is *partial correlation separable* (PC-separable) if it is of the form

$$Pen(\theta, \Delta) = \sum_i pen_{ii}(\theta_{ii}) + \sum_{i<j} pen_{ij}(\Delta_{ij}),$$

where $pen_{ii} : (0, \infty) \to \mathbb{R}$ and $pen_{ij} : [-1, 1] \to \mathbb{R}$ are non-decreasing in θ_{ii} and $|\Delta_{ij}|$ respectively, for all i and $i < j$.

A PC-separable penalty function is *symmetric* if $pen_{ii} = pen_{jj}$ for all (i, j) and, for all $i < j$, pen_{ij} does not depend on (i, j) and is symmetric about 0.

Note that Definition 5 includes formulations that do not penalise the diagonal entries, i.e. $pen_{ii}(\theta_{ii}) = 0$. Note also that the L_0 and logarithmic penalties are PC-separable since $\theta_{ij} = 0$ if and only if $\Delta_{ij} = 0$ and $\log(|\theta_{ij}|) = \log(|\Delta_{ij}|) + \frac{1}{2}\log(\theta_{ii}) + \frac{1}{2}\log(\theta_{jj})$.

The PC-GLASSO can be considered the symmetric PC-separable counterpart to the GLASSO applying the L_1 norm to the partial correlations $pen_{ij}(\Delta_{ij}) = n\rho|\Delta_{ij}|$. On the diagonal entries a logarithmic penalty is applied $pen_{ii}(\theta_{ii}) = 2\log(\theta_{ii})$ - the motivation for this will be discussed in Sections 2.5 and 2.8. The penalised likelihood function, after removing constants, is given by

$$\log(\det(\Delta)) + \left(1 - \frac{4}{n}\right)\sum_i \log(\theta_{ii}) - \text{tr}\left(S\theta^{\frac{1}{2}}\Delta\theta^{\frac{1}{2}}\right) - \rho\sum_{i\neq j}|\Delta_{ij}|. \tag{2.5}$$

An important consideration for the PC-GLASSO is the choice of regularisation parameter ρ. As introduced in Section 1.4.1, one may consider a sequence of parameters $0 = \rho_0 < \rho_1 < \cdots < \rho_m$ and calculate the PC-GLASSO estimate for each of these parameters $\hat{\Theta}_{\rho_0}, \hat{\Theta}_{\rho_1}, \ldots, \hat{\Theta}_{\rho_m}$. The regularistion parameter is then selected by choosing the estimate that maximises some chosen criterion. In Section 2.10 we use the Bayesian information criterion (BIC), which selects the parameter minimising

$$\mathrm{BIC}(\hat{\Theta}_{\rho_i}, S) = \log(n) k_{\hat{\Theta}_{\rho_i}} - 2l(\hat{\Theta}_{\rho_i} \mid S), \tag{2.6}$$

where $k_{\hat{\Theta}_{\rho_i}}$ is the number of edges in the graphical model given by $\hat{\Theta}_{\rho_i}$. This BIC criterion was suggested in this context by Yuan and Lin [2007] and has also been investigated by Lian [2011] and Gao et al. [2012].

There are some examples of penalty functions for Gaussian graphical models based on partial correlations. Ha and Sun [2014] utilised a ridge penalty. The space method of Peng et al. [2009], similarly to PC-GLASSO, uses an L_1 penalty on the partial correlations, but in combination with a function other than the log-likelihood. Azose and Raftery [2018] introduced a separable prior on the *marginal* correlations. They argued that a key benefit of their prior is the ability to specify beliefs about correlations. A similar argument can be made for PC-separable priors, introduced in Chapter 3, allowing one to specify prior beliefs on partial correlations.

2.3 Data standardisation

It is a common practise when using Gaussian data to standardise the data before applying any statistical methods. In this section we will detail two ways in which this might be done.

The most common form of standardisation ensures that the sample marginal variances are all equal to 1, or equivalently that S has unit diagonal. We will refer to this form of standardisation as *standardising by S*. Here the sample correlation matrix

$$R = \mathrm{diag}(S)^{-1/2}\, S\, \mathrm{diag}(S)^{-1/2},$$

where $\mathrm{diag}(S)$ is the $p \times p$ diagonal matrix with diagonal entries equal to those of S, is substituted for the matrix S. To estimate the original (unstandardised) Θ, one can simply rescale the estimate obtained when using R. That is, if $\hat{\Theta}$ is an estimator, then

$$\mathrm{diag}(S)^{-1/2}\, \hat{\Theta}(R)\, \mathrm{diag}(S)^{-1/2} \tag{2.7}$$

can be considered an estimate for Θ.

One reason why standardising by S is common is because $\mathrm{diag}(S)$ is a good estimator for $\mathrm{diag}(\Sigma)$. In particular

$$(n-1)S \sim \mathcal{W}_p(\Sigma, n-1),$$

where $\mathcal{W}_p$ denotes the p dimensional Wishart distribution, and so S is an unbiased estimator for Σ. Furthermore, the marginal distributions of the diagonal entries of S satisfy

$$\frac{n-1}{\sigma_{ii}} S_{ii} \sim \mathcal{X}^2_{n-1},$$

where $\mathcal{X}^2_{n-1}$ denotes the chi-squared distribution with $n-1$ degrees of freedom. It follows that S_{ii} is an unbiased estimator of σ_{ii} with mode at $\frac{n-3}{n-1}\sigma_{ii}$ and variance $\frac{2}{n-1}\sigma^2_{ii}$ which does not depend on the dimension p. In particular, the probability

$$\mathbb{P}\left(S_{ii} \in (0.5\sigma_{ii}, 1.5\sigma_{ii})\right)$$

is equal to 0.734 for $n = 10$ and 0.999 for $n = 100$.

It is a therefore a justifiable assumption, particularly for large n, that under the standardised data the true underlying variances are all approximately equal to 1. In other words, the rescaled covariance matrix

$$\mathrm{diag}(S)^{-1/2}\, \Sigma\, \mathrm{diag}(S)^{-1/2}$$

has approximately unit diagonal entries.

An alternative form of standardisation is to make the sample partial variances all equal to 1, or equivalently to ensure that S^{-1} has unit diagonal. We refer to this form of standardisation as *standardising by* S^{-1}. This standardisation is achieved by considering

$$R_{-1} = \mathrm{diag}(S^{-1})^{1/2}\, S\, \mathrm{diag}(S^{-1})^{1/2}.$$

This time an estimate for the orginal Θ can be obtained via

$$\mathrm{diag}(S^{-1})^{1/2}\, \hat{\Theta}(R_{-1})\, \mathrm{diag}(S^{-1})^{1/2} \tag{2.8}$$

Standardising by S^{-1} is a less popular form of standardisation, in part due to the poor properties of S^{-1} as an estimator of Θ. In particular,

$$\frac{1}{n-1} S^{-1} \sim \mathcal{W}_p^{-1}(\Theta, n-1)$$

where $\mathcal{W}_p^{-1}$ denotes the p dimensional Inverse-Wishart distribution. Hence the expectation of S^{-1} is

$$\mathbb{E}[S^{-1}] = \frac{n-1}{n-p-2}\Theta$$

and so is a biased estimator for Θ. The marginal distributions of the diagonal entries of S^{-1} satisfy

$$\frac{1}{n-1}(S^{-1})_{ii} \sim \mathrm{IG}\left(\frac{1}{2}(n-p-2), \frac{1}{2}\theta_{ii}\right)$$

where IG denotes the inverse Gamma distribution. It follows that $(S^{-1})_{ii}$ is a biased estimate of θ_{ii} with mean $\frac{n-1}{n-p-4}\theta_{ii}$, mode $\frac{n-1}{n-p}\theta_{ii}$ and variance $\frac{2(n-1)^2}{(n-p-4)^2(n-p-6)}\theta_{ii}^2$. Note that both the bias and variance increase with the dimension p.

From this it follows that the diagonal entries of S^{-1} can be far from the diagonal entries of Θ, particularly when the dimension p is large. It is therefore unreasonable to assume that under the standardised data the true underlying partial variances are all approximately equal to 1. In particular, the diagonal entries of the rescaled precision matrix

$$\mathrm{diag}(S^{-1})^{-1/2}\,\Theta\,\mathrm{diag}(S^{-1})^{-1/2}$$

may be highly dispersed and far from 1. In other words, while standardising in terms of S^{-1} ensures that the sample partial variances are all equal to 1, this does not imply that the true partial variances are approximately equal to 1, or even that they are approximately equal to each other.

Furthermore, the situation gets worse in big data settings when $p > n$, where S is not invertible. Instead, a generalised inverse, such as the Moore-Penrose inverse, may be used. However, in this case the properties of the inverse as an estimate for Θ deteriorate further. More details on the Moore-Penrose inverse can be found in, for example, Cook and Forzani [2011]. For these reasons we will primarily focus on standardising by S for the remainder of this chapter.

As a final comment of this section, we note that standardisation of the data is not as innocuous as it may first seem. In particular, to standardise the data one must multiply by some function of the data. It follows that the distribution of the standardised data is no longer Gaussian, and that the sample correlation matrix R is not Wishart [Kollo and Ruul, 2003]. One may argue that it is therefore preferable to work with unstandardised data where possible.

2.4 A 2×2 comparison

Within linear regression settings it is common to gain insight into the shrinkage affects of a penalty function by considering the orthogonal design matrix case. In this case the likelihood is separable in each of the regression coefficients thus giving a closed form solution, with the estimated regression coefficients being independent of one another. In the Gaussian graphical model setting of this thesis, one may gain similar insights by considering the $p = 2$ case. In this section we compare the GLASSO and PC-GLASSO estimates for this $p = 2$ case. We begin with investigating GLASSO in the two cases where the data has been standardised by S and by S^{-1}. We then investigate PC-GLASSO for which the standardisation doesn't matter due to scale invariance, a property which will be introduced in Section 2.5.

2.4.1 GLASSO standardised by S

Consider the sample covariance matrix

$$S = \begin{pmatrix} 1 & x \\ x & 1 \end{pmatrix} \tag{2.9}$$

for $x \in [0, 1)$, which has inverse

$$S^{-1} = \begin{pmatrix} \frac{1}{1-x^2} & \frac{-x}{1-x^2} \\ \frac{-x}{1-x^2} & \frac{1}{1-x^2} \end{pmatrix}$$

which is the maximum likelihood estimate for the precision matrix. In this case the GLasso objective function can be written as

$$\log(\theta_{11}\theta_{22} - \theta_{12}^2) - \theta_{11} - \theta_{22} - 2x\theta_{12} - 2\rho|\theta_{12}| - \rho\theta_{11} - \rho\theta_{22}$$

and it can easily be shown that the GLasso estimate is

$$\theta_{12} = \begin{cases} 0, & 0 \le x < \rho \\ \frac{-(x-\rho)}{(1+\rho)^2-(x-\rho)^2}, & x > \rho \end{cases}$$

$$\theta_{11} = \theta_{22} = \begin{cases} \frac{1}{1+\rho}, & 0 \le x < \rho \\ \frac{1+\rho}{(1+\rho)^2-(x-\rho)^2}, & x > \rho \end{cases}$$

Note that this is equal to the MLE when the diagonal entries of S are replaced by $1 + \rho$ and the off-diagonal by $\max\{0, x - \rho\}$. Hence the penalty can be thought of as shrinking the off-diagonal entry of S towards 0.

We first consider a fixed penalty parameter $\rho = 0.1$ and allow x to vary between 0 and 1. Plots of the MLE vs the GLASSO estimate can be seen in Figure 2.2 for the partial correlations (left) and the diagonal entries (right). We see that similar to a LASSO style shrinkage is applied to the partial correlations. A large amount of shrinkage is applied to the diagonal entries, particularly for large x, with the GLASSO estimate remaining less than 3 even as the MLE goes above 100.

Next we consider fixed $x = 0.5$ and allow the penalty ρ to vary between 0 and 1. In Figure 2.3 we see the GLASSO estimates plotted against the penalty parameter for the partial correlations (left) and diagonal entries (right). Note that GLASSO shrinks both the partial correlations and diagonal entries to zero at a super-linear rate as the penalty parameter increases.

2.4.2 GLASSO standardised by S^{-1}

We now consider the inverse sample covariance matrix

$$S^{-1} = \begin{pmatrix} 1 & -x \\ -x & 1 \end{pmatrix}.$$

for $x \in [0, 1)$, which has inverse

$$S = \begin{pmatrix} \frac{1}{1-x^2} & \frac{x}{1-x^2} \\ \frac{x}{1-x^2} & \frac{1}{1-x^2} \end{pmatrix}.$$

Once again, it can be shown that the GLASSO estimate is equivalent to replacing the diagonal entries of S by $\frac{1}{1-x^2} + \rho$ and the off-diagonals by $\max\{0, \frac{x}{1-x^2} - \rho\}$.

We now compare the MLE to the GLASSO estimate when $\rho = 0.1$ is fixed and x is allowed to vary between 0 and 1. In order to be comparable to the previous example, we multiply the MLE and GLASSO estimates by $\frac{1}{1-x^2}$. In the left panel of Figure 2.4 we see that the partial correlations receive LASSO style shrinkage for small values of x. However, as x increases, and the MLE becomes larger, the amount of shrinkage is reduced. Meanwhile, the right panel of Figure 2.4 shows that the diagonal entries receive far less shrinkage when the data is standardised by S^{-1}, in comparison to the results in Figure 2.2.

In Figure 2.5 we see that when $x = 0.5$ is fixed, the partial correlation estimates and diagonal entry estimates are shrunk towards zero at a super-linear rate. This is very similar to the results in Figure 2.3.

Figure 2.2: 2×2 example, standardised by S. The GLASSO estimate with penalty $\rho = 0.1$ of the partial correlation (left) and θ_{ii} (right) compared with the MLE as sample covariance x varies between 0 and 1. Dotted line is identity for reference.

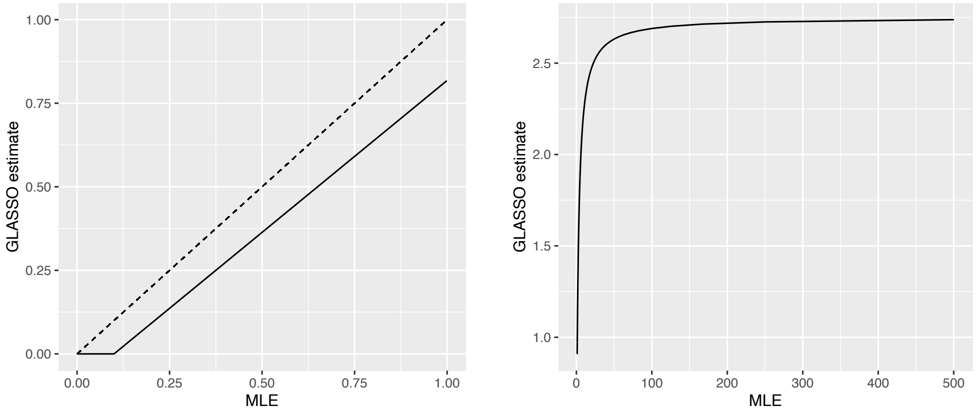
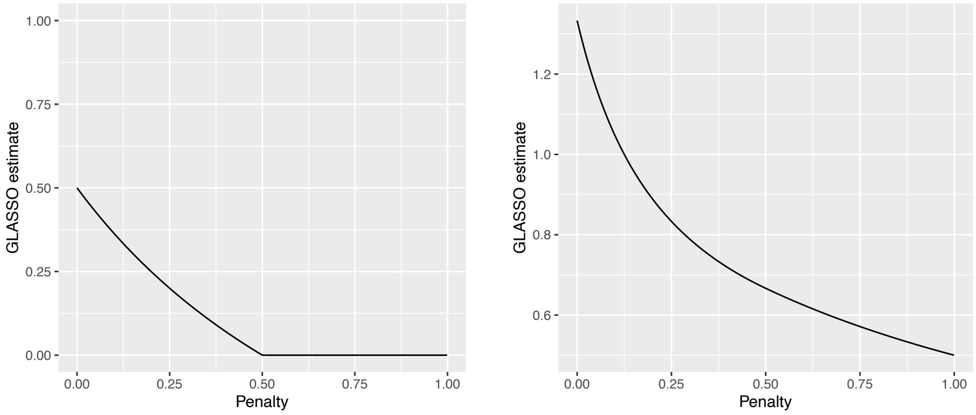


Figure 2.3: 2×2 example, standardised by S. The GLasso estimate with $x = 0.5$ of the partial correlation (left) and θ_{ii} (right) for different penalty parameter values.



2.4.3 PC-GLASSO

We now investigate the PC-GLASSO estimate in the $p = 2$ case. We will use the standardised sample covariance matrix (2.9) as in Section 2.4.1, however the results of Section 2.5 will show that the graphs produced in this section do not depend on the standardisation.

The PC-GLASSO objective function can be written as:

$$\log(1 - \Delta_{12}^2) + \left(1 - \frac{4}{n}\right)(\log(\theta_{11}) + \log(\theta_{22})) - \theta_{11} - \theta_{22} - 2x\sqrt{\theta_{11}\theta_{22}}\Delta_{12} - 2\rho|\Delta_{12}|.$$

The solution for Δ_{12} can be shown to be equal to 0 if $\left(1 - \frac{4}{n}\right) x < \rho$, and otherwise

Figure 2.4: 2×2 example, standardised by S^{-1}. The GLasso estimate with penalty $\rho = 0.1$ of the partial correlation (left) and θ_{ii} (right) compared with the MLE as sample partial covariance x varies between 0 and 1. Dotted line is identity for reference.

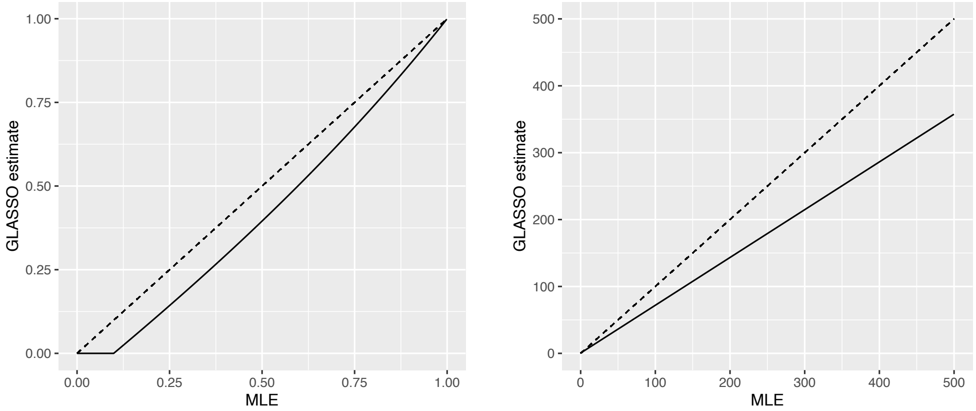
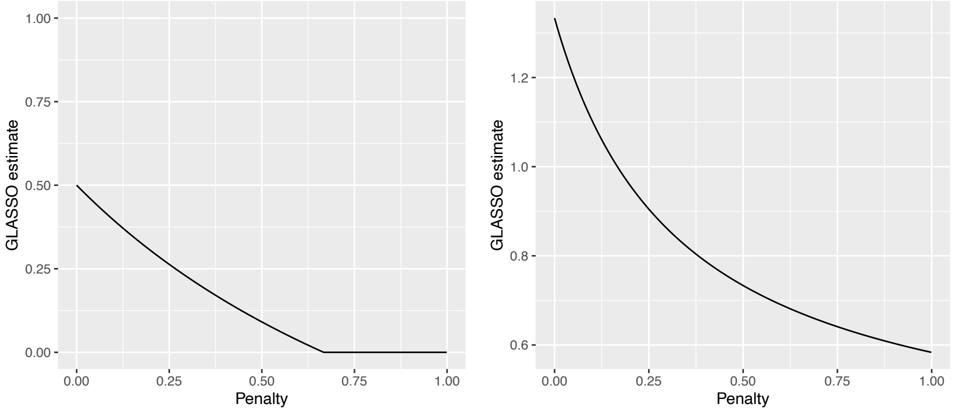


Figure 2.5: 2×2 example, standardised by S^{-1}. The GLasso estimate with $x = 0.5$ of the partial correlation (left) and θ_{ii} (right) for different penalty parameter values.



is the solution in $(-1, 0)$ of the cubic:

$$\rho x \Delta_{12}^3 + \left(\frac{4}{n}x + \rho\right)\Delta_{12}^2 + (1 - \rho x)\,\Delta_{12} + \left(1 - \frac{4}{n}\right)x - \rho = 0.$$

The solution for the diagonal entries is:

$$\theta_{11} = \theta_{22} = \frac{1 - \frac{4}{n}}{1 + x\Delta_{12}}.$$

Figure 2.6 compares the PC-GLASSO estimate to the MLE for fixed $\rho = 0.1$ as x varies between 0 and 1. Note that these look similar to those of Figure 2.4

when GLASSO is performed on data standardised by S^{-1}, although PC-GLASSO applies slightly less shrinkage on both the partial correlations and diagonal entries.

In Figure 2.7 we see the PC-GLASSO estimates for different penalty parameters ρ and fixed $x = 0.5$. Unlike GLASSO, we see that these are shrink towards 0 at a sub-linear rate.

We argue that these results are preferable to GLASSO because they apply less shrinkage to non-zero partial correlation estimates and to the diagonal entries, whilst still providing zero estimates. The goal of the penalty function is to induce sparsity in the estimated Θ, but when a partial correlation is not shrunk to 0 it is commonly thought of as preferable for the estimate to receive little shrinkage. This is the goal of non-convex penalty functions which aim to reduce the bias in non-zero estimates.

Furthermore, the results of PC-GLASSO do not depend on the standardisation. In this $p = 2$ example, GLASSO seems to provide better shrinkage when the data is standardised by S^{-1}. PC-GLASSO achieves even better shrinkage than this regarless of standardisation, in particular when data has not been standardised or the more common standardisation by S.

2.5 Scale invariance

A key property of graphical models is invariance to scalar multiplication. In the Gaussian case, if we consider the transformation DX for some fixed diagonal $p \times p$ matrix D with non-zero entries, then DX is also Gaussian with precision matrix

$$\Theta_D = D^{-1}\Theta D^{-1}. \tag{2.10}$$

In particular, the zero entries of Θ_D are identical to those of Θ.

We argue that it is desirable for an estimator of Θ to mirror the relationship in (2.10) under scalar multiplication of the data, a property we call *scale invariance*. We now show that, among regular penalty functions, only the L_0 and logarithmic penalties are scale invariant, whereas any PC-separable penalty with logarithmic diagonal penalty is scale invariant. We start by defining two notions of scale invariance related to the point estimate and to the recovered graphical structure.

Definition 6. An estimator $\hat{\Theta}$ is *scale invariant* if for any sample covariance matrix S and any diagonal $p \times p$ matrix D with non-zero diagonal entries,

$$\hat{\Theta}(DSD) = D^{-1}\hat{\Theta}(S)D^{-1}.$$

Figure 2.6: 2×2 example, PC-GLASSO. The PC-GLASSO estimate with $\rho = 0.1$ of the partial correlation (left) and θ_{ii} (right) compared with the MLE as x varies between 0 and 1. Dotted line is identity for reference.

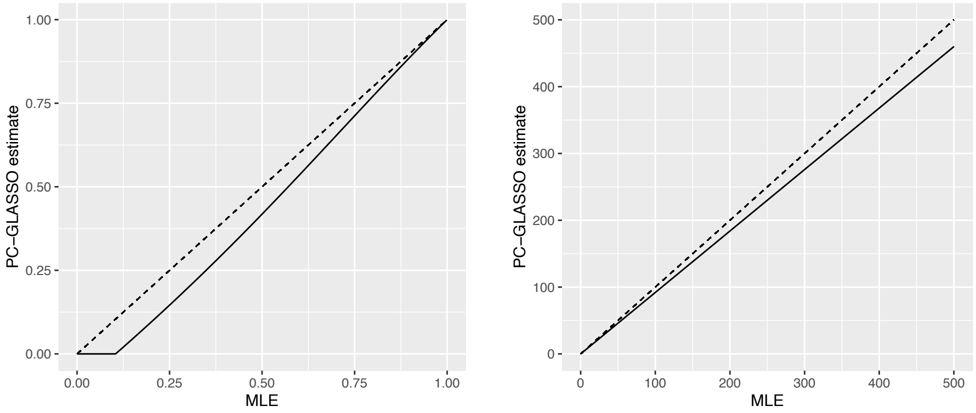
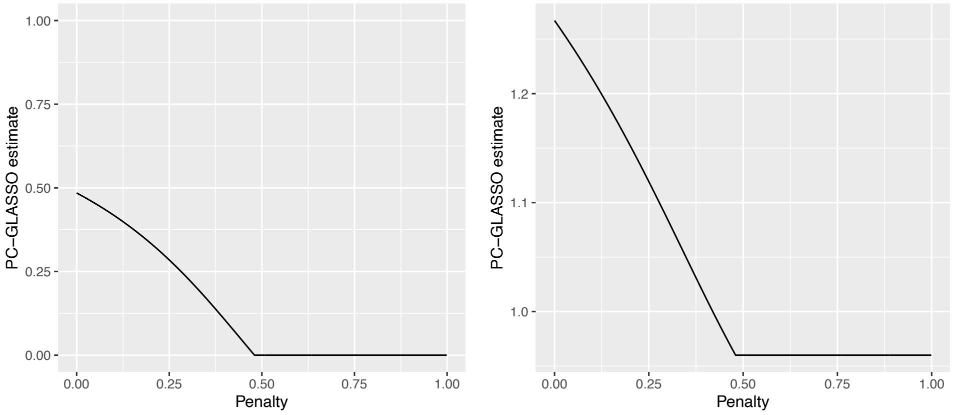


Figure 2.7: 2×2 example, PC-GLASSO. The PC-GLASSO estimate with $x = 0.5$ of the partial correlation (top) and θ_{ii} (bottom) for different penalty parameter values.



$\hat{\Theta}$ is *selection scale invariant* if $\hat{\Theta}(S)$ and $\hat{\Theta}(DSD)$ have identical zero entries for any S and D.

Scale invariance ensures that the estimate under the scaled data corresponds to that under the original data as in (2.10). In particular, the estimated partial correlations are identical in $\hat{\Theta}(S)$ and $\hat{\Theta}(DSD)$. Meanwhile selection scale invariance ensures that one recovers the same graphical structure under scalar multiplications. It is clear that scale invariance implies selection scale invariance.

We now present results on the scale invariance of different penalties.

Proposition 2. *Let $\hat{\Theta}$ be an estimator based on a regular penalty, and suppose that there exists a sample covariance matrix S such that $\hat{\Theta}(S)$ is not a diagonal matrix.*

Then $\hat{\Theta}$ is scale invariant if and only if pen_{ij} is either an L_0 or logarithmic penalty, and pen_{ii} is either a constant or a logarithmic penalty.

Proof. Let S be some sample covariance matrix for which $\hat{\Theta}(S)$ is not diagonal and D be some diagonal matrix with non-zero diagonal entries d_i, $i = 1, \ldots, p$. Suppose that $\hat{\Theta}$ is scale invariant. Let $\hat{\theta}_{ij} = \hat{\Theta}(S)_{ij}$ be some non-zero off-diagonal entry of $\hat{\Theta}(S)$, and $\tilde{\theta}_{ij} = \hat{\Theta}(DSD)_{ij}$ be the corresponding entry in $\hat{\Theta}(DSD)$. By scale invariance we must have $\tilde{\theta}_{ij} = \frac{\hat{\theta}_{ij}}{d_i d_j}$.

For these to maximise their corresponding penalised likelihoods, the derivatives of the penalised likelihood function (2.1) with respect to θ_{ij} must be equal to 0 at $\hat{\theta}_{ij}$ and $\tilde{\theta}_{ij}$ respectively (note that the derivative exists because Pen is regular and $\hat{\theta}_{ij} \neq 0, \tilde{\theta}_{ij} \neq 0$). Therefore

$$(\hat{\Theta}(S)^{-1})_{ij} - 2s_{ij} - \frac{4}{n} pen'_{ij}(\hat{\theta}_{ij}) = 0,$$

$$(\hat{\Theta}(DSD)^{-1})_{ij} - 2d_i d_j s_{ij} - \frac{4}{n} pen'_{ij}(\tilde{\theta}_{ij}) = d_i d_j \left((\hat{\Theta}(S)^{-1})_{ij} - 2s_{ij} \right) - \frac{4}{n} pen'_{ij} \left(\frac{\hat{\theta}_{ij}}{d_i d_j} \right)$$
$$= 0,$$

where we used that, since $\hat{\Theta}$ is scale invariant then $\hat{\Theta}(DSD) = D^{-1}\hat{\Theta}(S)D^{-1}$ and hence $(\hat{\Theta}(DSD)^{-1})_{ij} = (D\hat{\Theta}(S)^{-1}D)_{ij} = d_i d_j (\hat{\Theta}(S)^{-1})_{ij}$.

It follows that

$$pen'_{ij} \left(\frac{\hat{\theta}_{ij}}{d_i d_j} \right) = d_i d_j pen'_{ij}(\hat{\theta}_{ij}). \tag{2.11}$$

That is, for scale invariance to hold the penalty must satisfy $pen'_{ij}\left(\frac{\hat{\theta}_{ij}}{d}\right) = d pen'_{ij}(\hat{\theta}_{ij})$ for any $d \neq 0$. The latter requirement can only hold in two scenarios. First, there is the trivial scenario where $pen'_{ij}(\theta_{ij}) = 0$ for all $\theta_{ij} \neq 0$, that is pen_{ij} is an L_0 penalty.

Second, if $pen'_{ij}(\hat{\theta}_{ij}) = k \neq 0$, then $pen'_{ij}\left(\frac{\hat{\theta}_{ij}}{d}\right) = dk$. Treating $\hat{\theta}_{ij}$, and therefore also k, as fixed, we denote by $x = \frac{\hat{\theta}_{ij}}{d}$. Then we have $pen'_{ij}(x) = \frac{\hat{\theta}_{ij} k}{x}$. It follows that $pen_{ij}(x) = \hat{\theta}_{ij} k \log(|x|) + c$ for some constant c and $x \neq 0$, that is pen_{ij} is a logarithmic penalty.

This proves that for a regular penalty to be scale invariant it must have L_0 or logarithmic pen_{ij}. We now turn our attention to the diagonal penalty.

Let S be some diagonal covariance matrix, and D some diagonal matrix as before. Let $\hat{\theta}_{ii} = \hat{\Theta}(S)_{ii}$ and $\tilde{\theta}_{ii} = \hat{\Theta}(DSD)_{ii}$. By scale invariance we must have $\tilde{\theta}_{ij} = \frac{\hat{\theta}_{ij}}{d_i^2}$.

Since S is diagonal, it is easy to see that both $\hat{\Theta}(S)$ and $\hat{\Theta}(DSD)$ must also be diagonal, and that $\hat{\theta}_{ii}$ maximises the function:

$$\log(\theta_{ii}) - S_{ii}\theta_{ii} - \frac{2}{n} pen_{ii}(\theta_{ii}),$$

while $\tilde{\theta}_{ii}$ maximises the same function but with S_{ii} replaced by $d_i^2 S_{ii}$. It follows that the corresponding derivatives must both be equal to zero at $\hat{\theta}_{ii}$ and $\tilde{\theta}_{ii}$ respectively (Pen is regular so pen_{ii} is differentiable). Using this along with $\tilde{\theta}_{ij} = \frac{\hat{\theta}_{ij}}{d_i^2}$ we obtain:

$$pen'_{ii}\left(\frac{\hat{\theta}_{ii}}{d_i^2}\right) = d_i^2 pen'_{ii}\left(\hat{\theta}_{ii}\right).$$

As before, it follows that pen_{ii} must be either constant or logarithmic. This proves that for a regular penalty function to be scale invariant it must have either constant or logarithmic penalty on the diagonal entries.

To complete the proof we must show that such penalty functions (L_0 or logarithmic off-diagonal penalty and constant or logarithmic diagonal penalty) are always scale invariant. This follows from Proposition 3 since the L_0 and logarithmic penalties are also symmetric PC-separable.

□

It follows from Proposition 2 that the GLASSO, SCAD and MCP estimators are not scale invariant. Further, as illustrated in Figure 2.1 and the upcoming example these estimators are also not selection scale invariant. We conjecture that lack of selection scale invariance holds more widely for regular penalty functions, but settle with the counterexample for these three cases provided by Figure 2.1.

It should be noted that any penalised likelihood method can be made to be scale invariant by first standardising the data and then rescaling the estimate via (2.10). This is because the standardisation step removes the affect of the scalar multiplication - for example the sample correlation matrix R is invariant to scalar multiplication of the variables.

We present an example to further illustrate how scaling can affect the inferred conditional independence structure. Suppose we observe the inverse sample covariance matrix

$$S^{-1} = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.25 \\ 0 & 0.25 & 1 \end{pmatrix}$$

The left panel in Figure 2.8 shows the associated GLASSO estimates $\Theta^{\rho}_{\text{GLASSO}}(S)$. The right panel considers the situation where the data was given on a different scale, specifically the sample covariance is DSD where D has diagonal entries 1, 1 and 10, and provides the estimates $D\Theta^{\rho}_{\text{GLASSO}}(DSD)D$. The estimates set to zero, as well as their relative magnitudes, differ significantly depending on the scale of the data. We observed similar results for the SCAD and MCP penalties (not shown, for brevity).

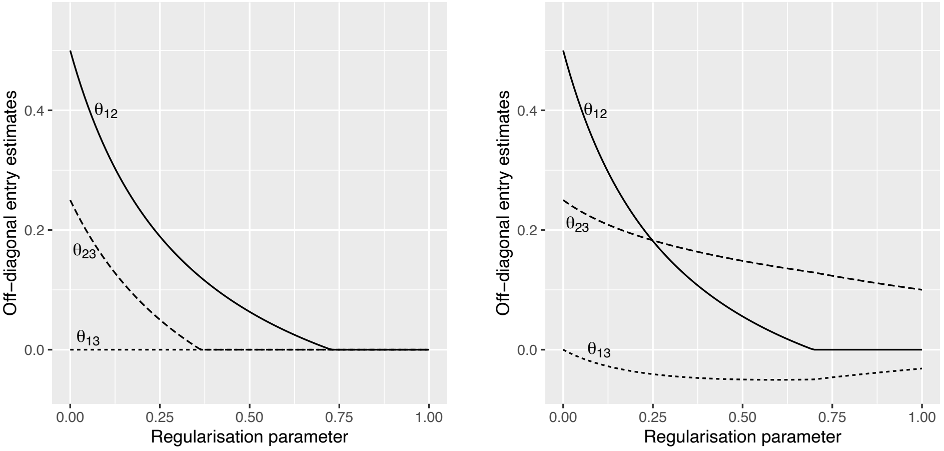


Figure 2.8: Estimated off-diagonal entries $\Theta^{\rho}_{\text{GLASSO}}(S)$ (left) and $D\Theta^{\rho}_{\text{GLASSO}}(DSD)D$ (right) for regularisation parameter $\rho \in [0, 1]$.

As shown in Proposition 2, the only scale invariant regular penalties are the L_0 and logarithmic penalties, both of which are also PC-separable. In fact scale invariance holds more widely in PC-separable penalties, from which it follows that PC-GLASSO is scale invariant.

Proposition 3. *Any estimator based on a symmetric PC-separable penalty with* $pen_{ii}(\theta_{ii}) = c\log(|\theta_{ii}|)$ *for some constant* $c \geq 0$ *is scale invariant.*

Proof. Let S be a sample covariance matrix and D be a diagonal matrix with non-zero entries d_i. Suppose that the estimate $\hat{\Theta}(S)$ decomposes as $\bar{\theta}^{1/2}\bar{\Delta}\bar{\theta}^{1/2}$ and that the estimate $\hat{\Theta}(DSD)$ decomposes as $\tilde{\theta}^{1/2}\tilde{\Delta}\tilde{\theta}^{1/2}$. To prove scale invariance we need that $\bar{\Delta} = \text{sign}(D)\tilde{\Delta}\text{sign}(D)$ and $\bar{\theta} = D^2\tilde{\theta}$.

Since $\hat{\Theta}(S)$ maximises the penalised likelihood at S, $\bar{\theta}, \bar{\Delta}$ must maximise

$$\log(\det(\Theta)) + \sum_i \left(\left(1 - \frac{2c}{n}\right) \log(\theta_{ii}) - s_{ii}\theta_{ii} \right) - \sum_{i \neq j} \left(s_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij} + \frac{2}{n} pen_{ij}(\Delta_{ij}) \right), \tag{2.12}$$

and similarly, $\tilde{\theta}, \tilde{\Delta}$ must maximise

$$\log(\det(\Theta)) + \sum_i \left(\left(1 - \frac{2c}{n}\right) \log(\theta_{ii}) - d_i^2 s_{ii}\theta_{ii} \right) - \sum_{i \neq j} \left(d_i d_j s_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij} + \frac{2}{n} pen_{ij}(\Delta_{ij}) \right). \tag{2.13}$$

By substituting $\theta'_{ii} = d_i^2\theta_{ii}$ and $\Delta'_{ij} = \text{sign}(d_i d_j)\Delta_{ij}$ into (2.13), and noting that pen_{ij} is symmetric about 0, we get

$$\begin{aligned} \log(\det(\Theta)) + \sum_i \left(\left(1 - \frac{2c}{n}\right) \left(\log(\theta'_{ii}) - \log(d_i^2)\right) - s_{ii}\theta'_{ii} \right) \\ - \sum_{i \neq j} \left(s_{ij}\sqrt{\theta'_{ii}\theta'_{jj}}\Delta'_{ij} + \frac{2}{n} pen_{ij}(\Delta'_{ij}) \right). \end{aligned} \tag{2.14}$$

Since $\log(d_i^2)$ is a constant, (2.14) is of the same form as (2.12) and they are maximised at the same point. Hence we have that $\bar{\Delta} = \text{sign}(D)\tilde{\Delta}\text{sign}(D)$ and $\bar{\theta} = D^2\tilde{\theta}$.

□

Proposition 3 states than any symmetric PC-separable is scale invariant, provided that the penalty function on the diagonal entries is logarithmic. Note that this also includes the case of no penalty on the diagonal entries by taking $c = 0$. The logarithmic penalty on the diagonal entries will be discussed further is Section 2.8.

2.6 Uneven penalisation

The previous section proved that regular penalty functions are not scale invariant and we saw in the examples at the beginning of this chapter that this lack of invariance, at least in the case of GLASSO, is not benign - rescaling the variables has a significant effect on both estimation of Θ and graphical model selection. In those examples, GLASSO performed best when the data generating θ_{ii} were all equal to 1. In this section we will provide some informal insights into why this might be the case more generally for both GLASSO and regular penalty functions. The following section will then attempt to formalise these ideas in a logical criterion we call exchangeable inference.

After removing constants, a penalised likelihood function with a regular penalty can be written as

$$\frac{n}{2}\left(\log(\det(\Theta)) - \operatorname{tr}(S\Theta)\right) - Pen(\Theta)$$

which is proportional to

$$\begin{aligned}&\log(\det(\Theta)) - \operatorname{tr}(S\Theta) - \frac{2}{n}Pen(\Theta)\\ &\qquad = \log(\det(\Theta)) - \sum_{i}\left(S_{ii}\theta_{ii} - \frac{2}{n}pen_{ii}(\theta_{ii})\right) - \sum_{i<j}\left(2S_{ij}\theta_{ij} - \frac{2}{n}pen_{ij}(\theta_{ij})\right).\end{aligned}$$

To better understand the penalised likelihood estimate, we will consider maximisation of this function ignoring the log determinant term. Although the log determinant is important to the maximisation problem, it does complicate the issue with the maximum not generally being analytically available. By not considering the log determinant we obtain a proxy to the maximum which is more generally analytically available. Furthermore, the log determinant term tends to act a a regulator for positive definiteness, with matrices close to the boundary of positive definiteness having a large negative value but with the log determinant being relatively constant on the interior of the space.

Ignoring the log determinant term, the value of θ_{ij} that maximises this function is determined solely by S_{ij}. In particular, if, without loss of generality, we suppose $pen_{ij}(0) = 0$, then this function will be maximised at $\theta_{ij} = 0$ if and only if $\frac{1}{n}pen_{ij}(\theta_{ij}) > S_{ij}\theta_{ij}$ for all $\theta_{ij} \neq 0$. Clearly the penalised likelihood will therefore generally select edges associated to larger S_{ij} in absolute value.

However, the magnitude of S_{ij} is strongly influenced by the scale of the variables and not on the strength of dependence between them. This is because $S_{ij} = \sqrt{S_{ii}S_{jj}}R_{ij}$ where R_{ij} is the sample correlation. While $|R_{ij}|< 1$ is bounded, S_{ii}, S_{jj} are unbounded. Hence the magnitude of S_{ij} is mostly determined by the sample variances. The penalised likelihood will therefore tend to select edges related to high variance variables.

This seems to suggest that regular penalties might perform best when all variables have equal variance (i.e. the diagonal entries of Σ are all equal). Although this is an unreasonable assumption in any real data set, this can be approximately achieved by standardising the data by S. However, as seen in the example at the beginning of this chapter, the performance of GLASSO can be poor even when the data has been standardised by S.

An alternative viewpoint is to consider two equal partial correlations $\Delta_{ij} = \Delta_{i'j'}$ but with diagonal entries $\theta_{ii}\theta_{jj} < \theta_{i'i'}\theta_{j'j'}$ so that $\theta_{ij} < \theta_{i'j'}$. Hence under a regular penalty $pen_{ij}(\theta_{ij}) < pen_{ij}(\theta_{i'j'})$. This seems to suggest that a regular penalty would favour edges associated to small diagonal entries of Θ and so might perform best when these are all equal. This was certainly the case in the example at the beginning of this chapter.

Again, it is unreasonable to assume that the diagonal entries of Θ are equal in any real data set. Furthermore, as discussed in Section 2.3, this assumption is not necessarily reasonable even after standardising the data by S^{-1}. Standardising by S^{-1} was favoured by Yuan and Lin [2007] stating that this seems more natural than standardising by S since we are estimating the precision matrix. They also claim that there is little difference in performance when standardising in either way, although in our experience this is not the case in certain settings.

2.7 Exchangeable inference

In this section we attempt to formalise the arguments of the previous section providing a logical criterion, which we call exchangeable inference, which is only satisfied for regular penalties when the data has been standardised.

The simplest situation of exchangeable inference occurs when the likelihood function is exchangeable in two or more Δ_{ij}'s, for example when two rows in the sample correlation matrix $R = \text{diag}(S)^{-1/2} S \text{diag}(S)^{-1/2}$ are equal (up to the necessary index permutations). In such a situation the likelihood provides the same information on these Δ_{ij}'s, hence it seems desirable to obtain the same inference for all of them. If the log-likelihood is exchangeable in some parameters, then any symmetric PC-separable penalty and prior trivially leads to exchangeable inference on those parameters. Yet, as illustrated in our example below, regular penalties can lead to significantly different inference (unless one standardises the data).

Example Consider a $p = 4$ setting where the data-generating truth follows a star graph, featuring an edge between $X^{(1)}$ and each of $X^{(2)}, X^{(3)}, X^{(4)}$, and no other edges. Specifically, suppose that truly $\theta_{11} = \theta_{22} = \theta_{44} = 1$, $\theta_{33} = 4$, $\theta_{12} = \theta_{14} = -0.5$ and $\theta_{13} = -1$, so that the data-generating partial correlations are $\Delta_{12} = \Delta_{13} = \Delta_{14} = 0.5$, and $\Delta_{ij} = 0$ for all remaining (i, j). Consider an ideal scenario where the sample covariance S matches the data-generating truth. That is,

$$S^{-1} = \begin{pmatrix} 1 & -0.5 & -1 & -0.5 \\ -0.5 & 1 & 0 & 0 \\ -1 & 0 & 4 & 0 \\ -0.5 & 0 & 0 & 1 \end{pmatrix}; \quad S = \begin{pmatrix} 4 & 2 & 1 & 2 \\ 2 & 2 & 0.5 & 1 \\ 1 & 0.5 & 0.5 & 0.5 \\ 2 & 1 & 0.5 & 2 \end{pmatrix};$$

$$R = \begin{pmatrix} 1 & 1/\sqrt{2} & 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & 1 & 0.5 & 0.5 \\ 1/\sqrt{2} & 0.5 & 1 & 0.5 \\ 1/\sqrt{2} & 0.5 & 0.5 & 1 \end{pmatrix}$$

In this example, the likelihood is exchangeable in $(\Delta_{12}, \Delta_{13}, \Delta_{14})$, hence it seems desirable that $\hat{\Delta}_{12} = \hat{\Delta}_{13} = \hat{\Delta}_{14}$. The estimates for the remaining Δ_{ij} should ideally be close to 0, their true value.

The left panel of Figure 2.9 shows the GLASSO path for the partial correlations. The estimate for Δ_{13} is fairly different than for Δ_{12} and Δ_{14}, and so is the range of ρ's for which they are set to 0. Note however that the estimates for the remaining Δ_{ij}'s are close to 0. To address this issue, one may note that the diagonal of S is not equal to 1. Indeed, if one standardises the data by S, so that the sample covariance is equal to R, one obtains the center panel of Figure 2.9. Now $\hat{\Delta}_{12} = \hat{\Delta}_{13} = \hat{\Delta}_{14}$ for any regularisation parameter ρ, as we argued is desirable. However, the estimates for truly zero parameters are somewhat magnified for $\rho \in [0.05, 0.35]$.

The PC-GLASSO estimates (on either the original or standardised data, due to scale invariance) in the right panel of Figure 2.9 satisfy $\hat{\Delta}_{12} = \hat{\Delta}_{13} = \hat{\Delta}_{14}$, and the truly zero parameters are clearly distinguished.

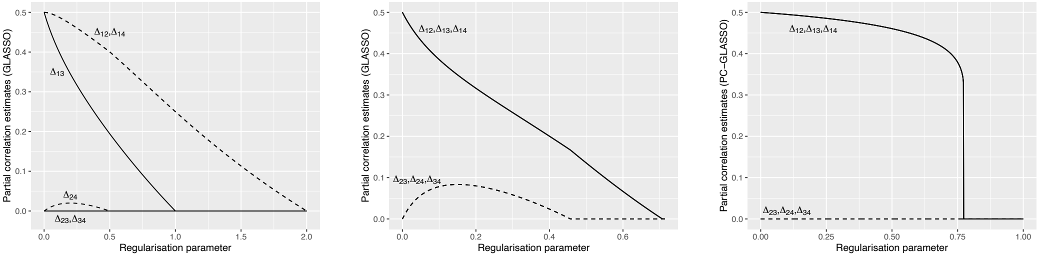


Figure 2.9: Partial correlation regularisation paths in $p = 4$ star graph example for GLASSO on the original S (left), standardised S (center) and PC-GLASSO (right).

We now extend this idea by constructing a situation where, given some values of the remaining parameters, the log-likelihood is symmetric in two partial correlations. Note that since the log-likelihood is concave, this implies that the two partial correlations are equal in the MLE. We argue that any penalised likelihood should match this symmetry so that inference on the two partial correlations is exchangeable.

Suppose the value of an estimator $\hat{\theta} = \text{diag}(\hat{\Theta})$ and all the entries in $\hat{\Delta}$ are given, except for a pair of partial correlations $(\Delta_{k_1k_2}, \Delta_{k_1k_3})$, for some indexes $k_1, k_2, k_3 \in \{1, \ldots, p\}$. Suppose that S, and the given elements in $\hat{\Delta}$ and $\hat{\theta}$ satisfy the following conditions:

(C1) $S_{k_1k_2} = cS_{k_1k_3}$ for some $c > 0$.

(C2) $\hat{\theta}_{k_2}^{-1/2} = c\hat{\theta}_{k_3}^{-1/2}$ for the same $c > 0$.

(C3) $\hat{\Delta}_{k_2j} = \hat{\Delta}_{k_3j}$ for all $j \notin \{k_1, k_2, k_3\}$.

Proposition 4. *Under conditions (C1)-(C3) the likelihood function is symmetric in* $(\Delta_{k_1k_2}, \Delta_{k_1k_3})$.

Proof. Without loss of generality suppose that the variable indexes are $k_1 = 1$, $k_2 = 2$ and $k_3 = 3$. The MLE maximises the function

$$\log(\det(\Theta)) - \text{tr}(S\Theta) = \log(\det(\theta^{1/2}\Delta\theta^{1/2})) - \text{tr}(S\theta^{1/2}\Delta\theta^{1/2}).$$

Consider this as a function $h(\Delta_{12}, \Delta_{13})$ that only depends on $(\Delta_{12}, \Delta_{13})$, given a value of the remaining parameters $\hat{\theta}$ and $\hat{\Delta}_{ij}$ for $(i, j) \notin \{(1, 2), (1, 3)\}$ satisfying (C1)-(C3).

We shall show that the two terms $\log\det(\Theta)$ and $\text{tr}(S\Theta)$ are symmetric in $(\Delta_{12}, \Delta_{13})$, when (C1)-(C3) hold. Using straightforward algebra gives that

$$\text{tr}(S\Theta) = \text{tr}(S\theta^{1/2}\Delta\theta^{1/2}) = 2s_{12}\theta_{11}^{1/2}\theta_{22}^{1/2}\Delta_{12} + 2s_{13}\theta_{11}^{1/2}\theta_{13}^{1/2}\Delta_{13} + c$$

where c does not depend on $(\Delta_{12}, \Delta_{13})$. Plugging in $\hat{\theta}$ and $\hat{\Delta}_{ij}$ into this expresion and using (C1) gives that is it equal to

$$2\hat{\theta}_{11}^{1/2}s_{12}\hat{\theta}_{22}^{1/2}(\Delta_{12} + \Delta_{13}) + c, \tag{2.15}$$

which is symmetric in $(\Delta_{12}, \Delta_{13})$.

Consider now $\det(\Theta)$. Using basic properties of the matrix determinant,

$$\det(\Theta) = \det(\Delta)\prod_{j=1}^{p}\theta_{jj} = |\Delta_{11} - \Delta_{2:p,1}\Delta_{2:p,2:p}^{-1}\Delta_{1,2:p}||\Delta_{2:p,2:p}|\prod_{j=1}^{p}\theta_{jj},$$

where $\Delta_{i:j,k:l}$ is the submatrix obtained by taking rows $i, i+1, \ldots, j$ and columns $k, k+1, \ldots, l$ from Δ. Since $\hat{\theta}$, $\hat{\Delta}_{2:p,2:p}$, and $\hat{\Delta}_{1j}$ for $j \geq 4$ are given, it suffices to show that

$$(\Delta_{12}, \Delta_{13}, \hat{\Delta}_{14}, \ldots, \hat{\Delta}_{1p})\hat{\Delta}_{2:p,2:p}^{-1}(\Delta_{12}, \Delta_{13}, \hat{\Delta}_{14}, \ldots, \hat{\Delta}_{1p})^T \tag{2.16}$$

is symmetric in $(\Delta_{12}, \Delta_{13})$. To ease notation let $A = \hat{\Delta}_{2:p,2:p}^{-1}$. Note that under Condition (C3),

$$\hat{\Delta}_{2:p,2:p} = \begin{pmatrix} 1 & \hat{\Delta}_{23} & \hat{\Delta}_{24} & \ldots & \hat{\Delta}_{2p} \\ \hat{\Delta}_{23} & 1 & \hat{\Delta}_{24} & \ldots & \hat{\Delta}_{2p} \\ \ldots\hat{\Delta}_{2p} & \hat{\Delta}_{2p} & \hat{\Delta}_{4p} & \ldots & 1 \end{pmatrix}$$

and hence

$$\hat{\Delta}_{2:p,2:p}^{-1} = A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1p-1} \\ a_{12} & a_{11} & a_{13} & \ldots & a_{1p-1} \\ a_{13} & a_{23} & a_{33} & \ldots & a_{3p-1} \\ \ldots a_{1p-1} & a_{2p-1} & a_{3p-1} & \ldots & a_{p-1p-3} \end{pmatrix}.$$

That is, the first two rows in A are equal, up to permuting the first two elements in each row. Therefore, (2.16) is equal to

$$\begin{aligned} a_{11}\Delta_{12}^2 + a_{11}\Delta_{13}^2 + \sum_{j=3}^{p-1} a_{jj}\hat{\Delta}_{j+1j+1}^2 + 2a_{12}\Delta_{12}\Delta_{13} + 2\sum_{j=3}^{p-1} a_{1j}\Delta_{12}\hat{\Delta}_{1j+1} \\ +2\sum_{j=3}^{p-1} a_{1j}\Delta_{13}\hat{\Delta}_{1j+1} + 2\sum_{j=3}^{p}\sum_{k=j+1}^{p} a_{jk}\hat{\Delta}_{j+1k+1} \\ = a_{11}(\Delta_{12}^2 + \Delta_{13}^2) + 2a_{12}\Delta_{12}\Delta_{13} + 2(\Delta_{12} + \Delta_{13})\sum_{j=3}^{p-1} a_{1j}\hat{\Delta}_{1j+1} + c', \end{aligned}$$

where c' does not depend on $(\Delta_{12}, \Delta_{13})$, which is a symmetric function in $(\Delta_{12}, \Delta_{13})$, as we wished to prove.

□

Corollary 2. *Under conditions (C1)-(C3) any penalised likelihood with a symmetric PC-separable penalty is symmetric in* $(\Delta_{k_1k_2}, \Delta_{k_1k_3})$.

Proof. The proof follows immediately from the proof of Proposition 4, noting that $Pen(\theta, \Delta) = \sum_i pen_{ii}(\theta_{ii}) + \sum_{i \neq j} \text{pen}(\Delta_{ij})$ is symmetric in $(\Delta_{12}, \Delta_{13})$.

□

Corollary 3. *Under conditions (C1)-(C3) a penalised likelihood with a regular penalty, other than the* L_0 *or logarithmic, is symmetric in* $(\Delta_{k_1k_2}, \Delta_{k_1k_3})$ *if and only if* $\hat{\theta}_{k_2k_2} = \hat{\theta}_{k_3k_3}$.

Proof. From Proposition 4 the penalised likelihood is symmetric if and only if

$$pen_{k_1k_2}\left(\sqrt{\hat{\theta}_{k_1k_1}\hat{\theta}_{k_2k_2}}\Delta_{k_1k_2}\right) + pen_{k_1k_3}\left(\sqrt{\hat{\theta}_{k_1k_1}\hat{\theta}_{k_3k_3}}\Delta_{k_1k_3}\right)$$

is symmetric. Since Pen is regular, this only happens when $\hat{\theta}_{k_2k_2} = \hat{\theta}_{k_3k_3}$ or when pen_{ij} is either L_0 or logarithmic.

□

From Corollary 3, in order for a regular penalised likelihood to match the symmetry of the likelihood, it must estimate the appropriate diagonal entries to be equal. This will not be the case for most real data sets where there will likely be a large difference in the true values of the diagonal entries. This is another reason for standardising the data - under either standardisation and (C1)-(C3), the condition $\hat{\theta}_{k_2k_2} = \hat{\theta}_{k_3k_3}$ can be shown hold for the MLE.

2.8 Diagonal penalty

An important aspect of any separable or PC-separable penalty function on Θ is the penalty on the diagonal entries pen_{ii}. Good estimation of the diagonal entries is vital for both estimation of the partial correlations and graphical model selection due to their role as the inverse partial variances. The partial variance measures the variance in a variable after conditioning on the remaining variables. Hence, if one holds the marginal variance fixed, a large estimated partial variance implies weak dependence on the remaining variables, while a small partial variance implies strong dependence. From this it can be seen that underestimation of the diagonal entries may encourage a more sparse graph and vice versa.

The penalty on the diagonal entries is an aspect that is often overlooked in the study of regular penalties. A common approach is to simply use a function of the same form as the off-diagonal penalty. However, this approach seems suboptimal due to the differing objectives of the penalties. The aim of the penalty on the off-diagonal entries is to induce sparsity in order to obtain a more simple graphical model. On the other hand, the aim of the penalty on the diagonals is simply to improve estimation. The linear penalty of GLASSO $pen_{ii}(\theta_{ii}) = \rho\theta_{ii}$ will cause more shrinkage of the diagonal entries as the parameter ρ increases. The non-convex SCAD and MCP penalties apply more shrinkage to small diagonal entries than large ones. This will be seen in practise in the real data examples of Section 2.10.

An alternative approach used by Yuan and Lin [2007] is to simply not penalise the diagonals $pen_{ii}(\theta_{ii}) = 0$. While, in our opinion, this is preferable to the previous approach, it may still be improved upon. One reason we believe this is that S^{-1} is a biased estimate of Θ with

$$\mathbb{E}[(S^{-1})_{ii}] = \frac{n}{n-p-1}\theta_{ii}$$

and so some shrinkage is preferable.

In PCGLASSO we have opted for a logarithmic penalty, in part due to the property of scale invariance as detailed in Section 2.5. It is interesting to note why the logarithmic penalty is needed to obtain scale invariance. The derivative of $c\log(ax)$ with respect to x is equal to $\frac{c}{x}$ which does not depend on a. This means that the same shrinkage is applied to the diagonal entry after scalar multiplication - a property which only holds for the logarithmic (or zero) penalty.

Amongst the logarithmic penalty we have opted for a coefficient of 2. This is because amongst penalty functions of the form $c\log(x)$ for constant $c \geq 0$ on the precision, choosing $c = 2$ asymptotically minimises the mean squared error of the estimate of the precision in the univariate $p = 1$ case (detailed below). This is relevant in higher $p > 1$ dimensions since it relates to the diagonal entry estimate when all partial correlations are estimated to be equal to 0. That is, suppose we fix the estimates of all partial correlations to be equal to 0, $\hat{\Delta}_{ij} = 0$, and estimate the diagonal entries of Θ via penalised likelihood with a regular or PC-separable penalty. Then the estimates for the diagonal entries are equal to the estimates obtained by considering the p 1-dimensional problems. This gives some justification for the choice of $c = 2$, although it is an open question whether another choice may be optimal for $p > 1$.

Suppose we have n observations of $X \sim \mathrm{N}(\mu, \theta^{-1})$ with sample variance s.

Note that

$$(n-1)\theta s \sim \chi^2_{n-1},$$

and so

$$((n-1)\theta s)^{-1} \sim \mathrm{Inv} - \chi^2_{n-1}.$$

From this we get that

$$\mathbb{E}[s^{-1}] = \frac{n-1}{n-3}\theta,$$

$$\mathrm{Var}(s^{-1}) = \frac{2(n-1)^2}{(n-3)^2(n-5)}\theta^2.$$

Consider estimating θ via a penalised likelihood of the form

$$l(\theta \mid s) - c\log(\theta).$$

This can easily be shown to be maximised at

$$\hat{\theta} = \left(1 - \frac{2c}{n}\right) s^{-1}.$$

It follows that

$$\mathbb{E}[\hat{\theta}] = \left(1 - \frac{2c}{n}\right)\left(\frac{n-1}{n-3}\right)\theta,$$

$$\mathrm{Var}(\hat{\theta}) = \frac{2(1-\frac{2c}{n})^2(n-1)^2}{(n-3)^2(n-5)}\theta^2,$$

and so

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= \mathrm{Var}(\hat{\theta}) + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2 \\ &= \theta^2\left(\frac{2(1-\frac{2c}{n})^2(n-1)^2}{(n-3)^2(n-5)} + \left(\left(1-\frac{2c}{n}\right)\left(\frac{n-1}{n-3}\right) - 1\right)^2\right)\end{aligned}$$

It can be shown that this function is minimised at $c = \frac{2n}{n-1}$. Letting $n \to \infty$ we therefore get that the MSE is asymptotically minimised amongst logarithmic penalties by taking $c = 2$.

2.9 Computation

An important feature of GLASSO is its defining of a convex problem that significantly facilitates fast computation and its theoretical study. For example, Friedman et al. [2008] related GLASSO to a sequence of LASSO problems, see also Sustik and

Calderhead [2012] for improved algorithms. Computation for non-convex penalties such as SCAD and MCP poses a harder challenge, but the Local Linear Approximation of Zou and Li [2008] greatly facilitates this task, see also Fan et al. [2009]. The PC-GLASSO optimisation problem is non-convex. However it is conditionally convex given $\theta = \text{diag}(\Theta)$ so its geometry is still fairly simple and amenable to fast computation.

Proposition 5. *The penalised likelihood function (2.5) is concave in Δ, for any fixed value of θ.*

Proof. For a fixed θ, optimisation of the penalised likelihood function (2.5) is equivalent to optimisation of the following function

$$\log(\det(\Delta)) - \sum_{i \neq j} S_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij} - \rho \sum_{i \neq j} |\Delta_{ij}|.$$

The log-determinant function is known to be concave over the space of positive definite matrices. For fixed θ the second term is simply a sum of linear functions. The third term is simply a sum of clearly concave functions. Hence the objective function is a sum of concave functions and is therefore concave.

□

Proposition 5 opens the possibility to consider block-optimization algorithms, where $\hat{\theta}$ and $\hat{\Delta}$ are updated sequentially, to facilitate computation. In our examples, we took an even simpler strategy and used a coordinate descent algorithm. Despite its conceptual simplicity, the algorithm nevertheless requires the careful updating of each parameter to ensure positive definiteness of $\hat{\Delta}$. Details of the coordiante descent algorithm are given below. We have found on our test examples and simulations in this chapter that the algorithm typically converges in a few iterations. However, for higher dimensions the convergence may be significantly slower.

2.9.1 Coordinate descent algorithm

We now present the coordinate descent algorithm we used to calculate PC-GLASSO estimates in the simulated examples later in this chapter. Our aim is to find the values of Θ that maximise the objective function (2.5) for a sequence of penalty parameters $0 = \rho_0 < \rho_1 < \cdots < \rho_k$, i.e. the regularisation path. Algorithm 1, for which the coordinate descent algorithm 2 is embedded, specifies that the previous estimate related to ρ_{i-1} is used as a starting point for the coordinate descent for ρ_i. This ensures that the coordinate descent is initialised at a point close to the maximum and aids convergence. For $\rho_0 = 0$ the algorithm is initialised at S^{-1}, or at $(S + \alpha I)^{-1}$ where I is the identity matrix if $n < p$. Theoretically, the matrix $S + \alpha I$ is guaranteed to be invertible and positive definite for any α. Of course computationally the matrix may not be invertible if α is too small, and so in practise it should be chosen to be a small value for which computation of the inverse is still possible in the chosen programming language. An alternative option is

We also standardise the data by S, before returning the estimates to the original scale via (2.10). This has no effect on the estimated values due to the scale invariance of PC-GLASSO, however it helps with the numerics of the coordinate descent.

Algorithm 2 is a standard blockwise coordinate descent algorithm which randomly cycles through the entries of Δ and maximises the objective function with respect to $\Delta_{ij}, \Delta_{ji}, \theta_{ii}, \theta_{jj}$ while holding all other entries fixed. Once the algorithm has cycled through each of the entries of Δ exactly once, a stopping rule is tested. The stopping rule we choose is based on the increase in the value of the objective function brought about by the updates. If the increase in the objective function is less than a particular threshold then the algorithm is terminated and the current estimate is returned. Note that the threshold here is scaled by $q = \max\left\{\frac{2|\{\Delta_{ij}^{(0)} \neq 0 : i<j\}|}{p(p-1)}, \frac{2}{p(p-1)}\right\}$, the proportion of non-zero entries in the previous estimate $\Delta^{(0)}$. This is because once an entry is shrunk to zero, it is likely that it will remain zero in future estimates. Therefore, the number of entries that are actively being updated is proportional to q. If only a small number of entries are being actively updated then one would expect the increase in the objective function to be smaller. Hence, scaling the threshold by q helps to prevent the algorithm from terminating too early in situations where the current estimate is sparse.

Although no guarantees are made about the convergence of Algorithm 2, results in Patrascu and Necoara [2015] and Wright [2015] suggest that convergence towards a local maximum is guaranteed and give reasonable assurance of convergence towards the global maximum. Their results focus on a coordinate descent algorithm that cycles randomly through the indices *with* replacement and so are not directly applicable to Algorithm 2. However, we prefer cycling through the indices without replacement since this provides a more simple and clear stopping rule for the algorithm. Algorithm 2 assesses the convergence after updating each entry of Δ exactly once, so that the stopping rule at the end of each iteration is made on the same grounds. For an algorithm which selects indices with replacement it is less clear when to enact the stopping rule.

We also choose to randomise the order in which indices are cycled through at each iteration of the algorithm, rather than keeping a fixed randomisation over the iterations. This is to ensure that the maximisation is not driven by the ordering and help to avoid local maxima. A potential consequence of this would be the coordinate descent making large jumps between maxima leading to a non-smooth regularisation path. However, in our experience this is not generally the case with the regularisation path looking smooth, for example in Figure 2.1.

As a final note about Algorithm 2, Step 2 maximising (2.5) with respect to $\Delta_{ij}, \theta_{ii}, \theta_{jj}$ whilst all other variables are held fixed is non-trivial due to the non-smoothness of the objective function. Details of this maximisation problem, as well as an explanation of how the output of Algorithm 2 is guaranteed to be positive definite, provided that the starting point is positive definite, can be found in Appendix A.

```
Algorithm 1: PC-GLASSO regularisation path
    Input  : Sample covariance $S$, sequence of regularisation parameters
             $0 = \rho_0 < \rho_1 < \cdots < \rho_k$ and optimisation convergence
             threshold $\epsilon$.
    Output: Sequence of estimates $\Theta_0, \ldots, \Theta_k$ corresponding to $\rho_0, \ldots, \rho_k$.
      1. Standardise the sample covariance $\tilde{S} = \text{diag}(S)^{-1/2} S \text{diag}(S)^{-1/2}$.
      2. Run Algorithm 2 on $\tilde{S}$ for $\rho = 0$, with starting point $\Theta_0^{(0)} = \tilde{S}^{-1}$ (or
         $\Theta_0^{(0)} = (\tilde{S} + \alpha I)^{-1}$ for some $\alpha > 0$ if $n < p$), and threshold $\epsilon$ to obtain an
         estimate $\tilde{\Theta}_0$.
      3. For $i = 1, \ldots, k$, run Algorithm 2 on $\tilde{S}$ for penalty parameter $\rho = \rho_i$, with
         starting point $\Theta_i^{(0)} = \tilde{\Theta}_{(i-1)}$, and threshold $\epsilon$ to obtain an estimate $\tilde{\Theta}_i$.
      4. Return the sequence of estimates $\Theta_i = \text{diag}(S)^{-1/2} \tilde{\Theta}_i \text{diag}(S)^{-1/2}$ for
         $i = 0, 1, \ldots, k$.
```

2.10 Applications

In this section we will assess the performance of PC-GLASSO against other penalised likelihood methods in four simulation settings and two real data examples. In the simulation settings we will compare PC-GLASSO to only GLASSO, since they are directly comparable in the sense of using the same L_1 penalty structure. Results for the SCAD and MCP non-convex penalties for these simulation settings will be reported in Chapter 4 where we apply non-convex penalties. In the real-data settings we will compare PC-GLASSO with each of GLASSO, SCAD and MCP.

```
Algorithm 2: Blockwise coordinate descent
Input  : Sample covariance $S$ with unit diagonal, penalty parameter $\rho$,
         start point $\Theta^{(0)}$ and optimisation convergence threshold $\epsilon$.
Output: A matrix $\Theta$ providing a local maximum of (2.5) for penalty $\rho$.
  1. Let $\Theta^{(1)} = \Theta^{(0)}$ and decompose $\Theta^{(1)}$ to get $\theta^{(1)}$ and $\Delta^{(1)}$.
  2. Cycling randomly without replacement through the set of indices
     $\{(i,j) : i < j; i,j \in \{1,\dots,p\}\}$, let $\Delta_{ij}, \theta_{ii}, \theta_{jj}$ maximise
         $f(\Delta,\theta) = \log(\det(\Delta)) + \left(1 - \frac{4}{n}\right)\sum_i \log(\theta_{ii}) - \mathrm{tr}\left(S\theta^{1/2}\Delta\theta^{1/2}\right) - \rho\|\Delta\|_1,$
     subject to
         $\Delta_{k_1k_2} = \Delta^{(1)}_{k_1k_2}$, for all $(k_1,k_2) \neq (i,j)$,
         $\theta_{ii}, \theta_{jj} \geq 0$,
         $\theta_{kk} = \theta^{(1)}_{kk}$, for all $k \neq i,j$,
     and update $\Delta^{(1)}_{ij} = \Delta_{ij}$, $\Delta^{(1)}_{ji} = \Delta_{ji}$, $\theta^{(1)}_{ii} = \theta_{ii}$, $\theta^{(1)}_{jj} = \theta_{jj}$.
  3. Let $q = \max\left\{\frac{2|\{\Delta^{(0)}_{ij} \neq 0 : i<j\}|}{p(p-1)}, \frac{2}{p(p-1)}\right\}$ be the proportion of non-zero
     off-diagonal entries.
  4. If $f(\Delta^{(1)},\theta^{(1)}) - f(\Delta^{(0)},\theta^{(0)}) < q\epsilon$, set $\Delta = \Delta^{(1)}$, $\theta = \theta^{(1)}$ and return
     $\Theta = \theta^{1/2}\Delta\theta^{1/2}$. Otherwise, set $\Delta^{(0)} = \Delta^{(1)}$, $\theta^{(0)} = \theta^{(1)}$ and return to Step 2.
```

SCAD and MCP have an additional regularization parameter, which we set to the default proposed in Fan and Li [2001] and Zhang [2010] respectively. GLASSO was implemented using the R package **glasso** and SCAD and MCP using the package **GGMncv** (see Williams [2020]).

2.10.1 Simulations

We considered four simulation scenarios with Gaussian data, truly zero mean and precision matrix Θ with unit diagonal and off-diagonal entries as follows.

Scenario 1: The star graph - $\theta_{ij} = \begin{cases} -\frac{1}{\sqrt{p}}, & i = 1 \text{ or } j = 1 \\ 0, & \text{otherwise} \end{cases}$

Scenario 2: The hub graph - Partition variables into 4 groups of equal size, with each group associated to a ‘hub’ variable i. For any $j \neq i$ in the same group as i we set $\theta_{ij} = \theta_{ji} = \frac{-2}{\sqrt{p}}$ and otherwise $\theta_{ij} = 0$.

Scenario 3: The AR2 model - $\theta_{ij} = \begin{cases} \frac{1}{2}, & j = i-1, i+1 \\ \frac{1}{4}, & j = i-2, i+2 \\ 0, & \text{otherwise} \end{cases}$

Scenario 4: A random graph - randomly select $\frac{3}{2}p$ of the θ_{ij} and set their values to be uniform on $[-1, -0.4] \cup [0.4, 1]$, and the remaining $\theta_{ij} = 0$. Calculate the sum of absolute values of off-diagonal entries for each column. Divide each off-diagonal entry by 1.1 times the corresponding column sum and average this rescaled matrix with its transpose to obtain a symmetric, positive definite matrix.

For each scenario we tested the following six methods.

M1. PC-GLASSO

M2. PC-GLASSO, but with zero diagonal penalty

M3. GLASSO on data standardised by S

M4. GLASSO on data standardised by S^{-1}

M5. GLASSO with no diagonal penalty on data standardised by S

M6. GLASSO with no diagonal penalty on data standardised by S^{-1}

For the remainder of this section we will often refer to these methods by their numbers. Method M2 is included to provide a direct comparison with methods M5 and M6 i.e. to examine the affect of placing the L_1 penalty on partial correlations rather than directly on the precision matrix. Comparing M1 with M2 then assesses the affect of the logarithmic penalty on the diagonal entries. For each method, a regularisation path is obtained for a large sequence of parameter values and a single estimate is selected via the BIC in (2.6).

For each setting a dimension size of $p = 20$ was used, sample sizes $n \in \{30, 100\}$ were considered and 100 independent simulations were performed. To assess estimation accuracy the Kullback–Leibler (KL) loss

$$\mathrm{KL}(\Theta, \hat{\Theta}) = -\log(\det(\hat{\Theta})) + \mathrm{tr}(\hat{\Theta}\Theta^{-1}) + \log(\det(\Theta)) - p$$

was used. To assess model selection accuracy the Matthews correlation coefficient (MCC) was used

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}},$$

where TP, TN, FP and FN stand for the number of true positives, true negatives, false positives and false negatives (respectively) and measure the ability to recover the true edges in the graph corresponding to Θ. The MCC combines specificity and sensitivity into a single assessment and ranges between -1 and 1, where 1 indicates perfect model selection. More information on the MCC including justification for it's use over other measures can be found in, for example, Chicco and Jurman [2020].

The results are summarised in Figure 2.10 showing the mean KL loss and MCC over the 100 simulations. Figure 2.11 shows the proportion of the 100 simulations in which each edge was selected for methods M1 and M3. More detailed results, including Frobenius norm (F-norm), sensitivity and specificity, and standard errors of the various metrics over the 100 simulations are in Tables 2.1-2.4 with the best score in each category given in bold. The F-norm sums the Euclidean distances between each of the entries of the estimate and the true Θ

$$||\Theta - \hat{\Theta}||_F = \sqrt{\sum_{i,j}(\theta_{ij} - \hat{\theta}_{ij})^2}.$$

This provides an alternative measure of estimation accuracy to the KL loss. Unlike the KL loss, it considers all entries of Θ equally and independently which is why we prefer the KL loss which takes into account the fact that Θ is a Gaussian precision matrix and utilises a similar form to the log-likelihood function. Sensitivity and specificity are defined as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

and are commonly used to assess model selection.

We begin by comparing the four GLASSO methods. In both the Star and Hub settings, methods M5 and M6 outperformed M3 and M4 respectively and so not penalising the diagonal entries seems to be preferable, as anticipated by Section 2.8. In the AR2 and Random settings the results are less clear cut, however not penalising the diagonal tended to offer fairly large improvements in terms of estimation. Methods M4 and M6 also outperformed M3 and M5 respectively in the Star and Hub settings, showing that standardising by S^{-1} rather than by S is important in these cases. A possible reason for this is the large range in node degrees in these examples - i.e. some nodes have many edges while some have few edges. In these cases, standardisation by S can result in additional penalisation of the true edges, for the reasons set out in Section 2.6. Again, this difference is less clear cut in the

AR2 and Random settings.

We now assess the impact of penalising partial correlations rather than θ_{ij} by comparing method M2 with methods M5 and M6. Method M2 achieves a higher MCC than both M5 and M6 across all simulation settings, other than the $n = 100$ random graph. M2 also has the lowest KL loss in all $n = 100$ settings, although M6 offers slight improvements in KL loss in the $n = 30$ settings. This demonstrates that penalising partial correlations can lead to improvements in model selection, as well as leading to faster convergence towards the true Θ for fixed p as n grows, when selecting the regularisation parameter by BIC.

Now to investigate the affect of the logarithmic penalty on the diagonal entries, we compare methods M1 and M2. Method M1 tends to provide improvements in estimation for $n = 30$, but a negligible difference for $n = 100$ and in model selection. This shows that the logarithmic penalty on the diagonal entries can provide improvements in estimation for small sample sizes in comparison to no penalty.

Overall, PC-GLASSO offers significant improvements over GLASSO in terms of both estimation and model selection in both the Star and Hub settings. This demonstrates that penalising partial correlations is particularly beneficial when there is a large range of node degrees in the true underlying graph. There are also large improvements in estimation for small $n = 30$ sample sizes in the AR2 and Random graph settings, while model selection and estimation in the $n = 100$ case are very similar in performance throughout all methods in these settings. This demonstrates that when there is a small range in node degrees, PC-GLASSO still performs at least as well as GLASSO, and can also offer improvements for small sample sizes.

Figure 2.11, showing the proportion of the 100 simulations in which each edge was selected, also illustrates that PCGLASSO generally selected sparser models than GLASSO (M3), particularly in the Star and Hub scenarios. This is a useful property in itself since more sparse graphs are generally preferable for interpretation and explanation.

Parameter selection via the EBIC in PC-GLASSO was also investigated within these simulation settings. Although in certain cases this did offer minor improvements in model selection measured by the MCC, in all scenarios the estimation accuracy, measured by both the F-Norm and KL loss, was significantly worse than when selecting the parameter via the BIC. For brevity we have omitted the results from this section.

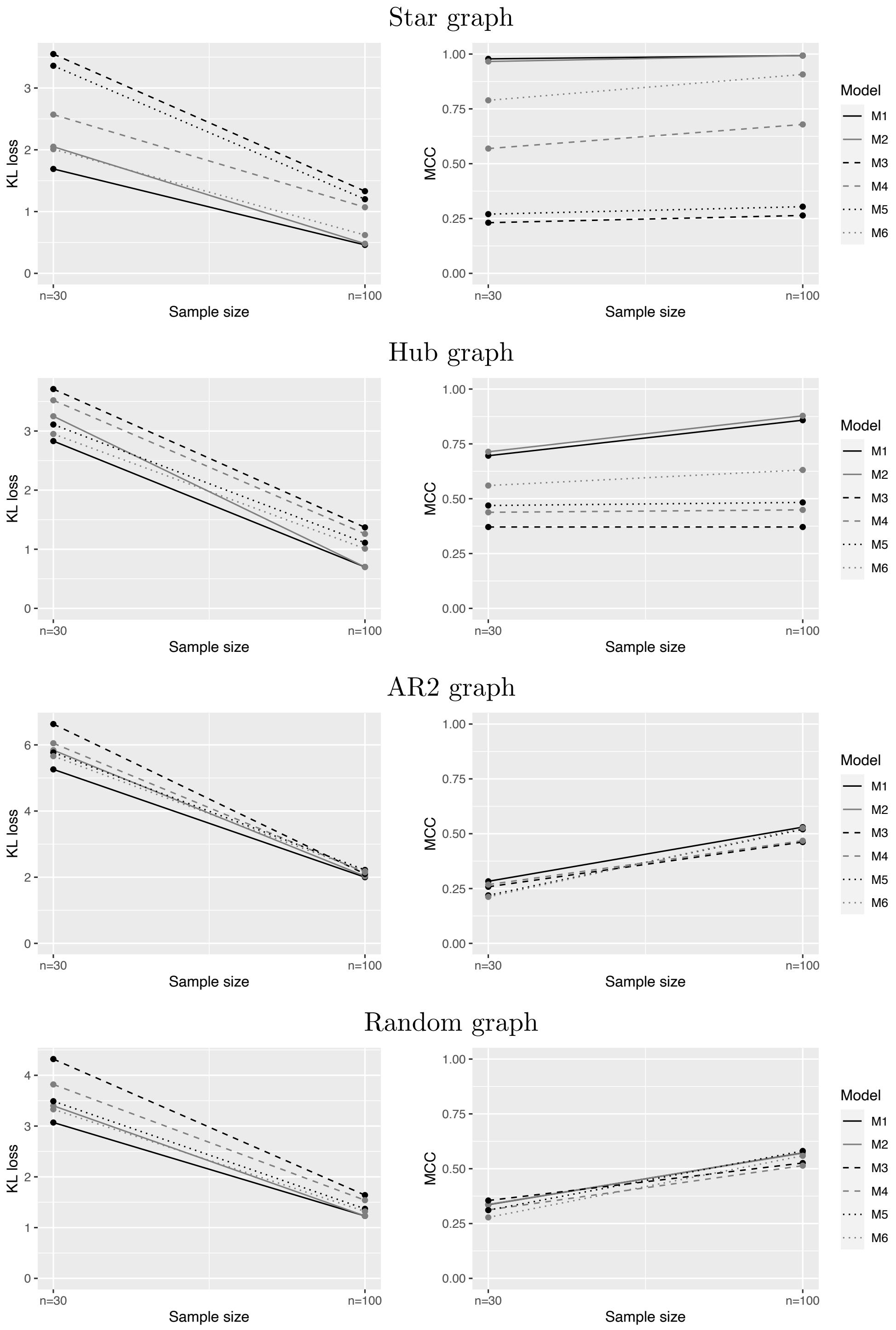


Figure 2.10: Kullback-Leibler loss (left) and MCC (right) in the four simulation settings

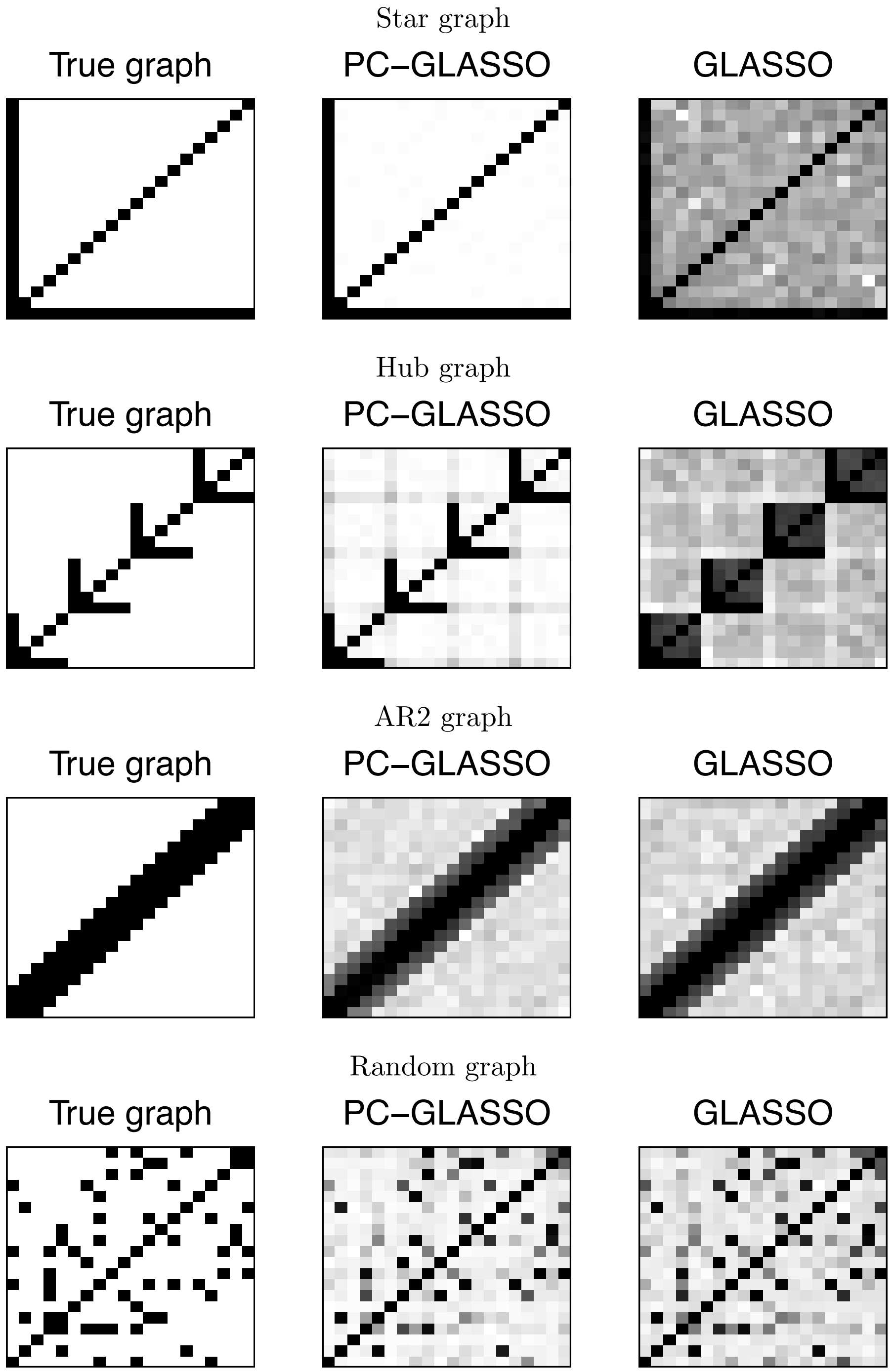


Figure 2.11: Proportion of simulations in which each edge was selected

2.10.2 Gene expression data

We now assess the predictive performance of PC-GLASSO in comparison to GLASSO (both with and without a diagonal penalty), SCAD and MCP in the gene expression data of Calon et al. [2012]. The data contains 262 observations of $p = 173$ genes related to colon cancer progression. We took $n = 200$ of the samples as training data, left the remaining 62 observations as test data, and assessed the predictive accuracy of each method by evaluating the log-likelihood on the test data. For each method this was performed for a long sequence of regularisation parameters chosen such that the null model (i.e. all partial correlations estimated to be 0) is selected for the largest regularisation parameter. SCAD and MCP have an additional regularization parameter, which we set to the default proposed in Fan and Li [2001] and Zhang [2010] respectively. For all methods, other than PC-GLASSO, we also standardise the data by either S or by S^{-1}, before rescaling estimates back to the original scale via (2.10).

Figure 2.12 plots the model size vs. test sample log-likelihood achieved by estimates in the regularisation path, and indicates the models chosen by the BIC and EBIC. We have restricted the view to model sizes less than 2000 since these contain all models chosen by BIC. The left panel compares PC-GLASSO to the other methods on data standardised by S. The right panel compares to the other methods on data standardised by S^{-1}.

First considering the left panel of Figure 2.12, PC-GLASSO clearly performs better than any other method when data is standardised by S. PC-GLASSO achieves a higher out of sample log-likelihood for model sizes smaller than 1900 in comparison to GLASSO with no diagonal penalty, and a higher log-likelihood score for all model sizes smaller than 2000 in comparison to the other methods. Furthermore, considering the estimates chosen by BIC and EBIC, PC-GLASSO achieves the highest log-likelihood score in both cases, even with a smaller model size than GLASSO in the case of BIC.

Now considering the right panel of Figure 2.12. It is clear that the other methods perform better when the data is standardised by S^{-1}. This may be due to a large range in node degrees, as in the Star graph example. Even so, PC-GLASSO still achieves the highest log-likelihood for all model sizes less than 1500, and has the highest log-likelihood at the BIC and EBIC estimates with a smaller or comparable model size to the remaining methods. This demonstrates the promise of PC-GLASSO - it shows that the method can achieve better prediction than the remaining methods with a simpler model.

One interesting thing to note in this example is the relatively poor performance of the non-concave SCAD and MCP penalties. When data is standardised by S they both achieve significantly lower log-likelihood than GLASSO for both large and very small model sizes, while when data is standardised by S^{-1} they perform almost identically to GLASSO - the method that they were designed to improve upon. We conjecture that the reason for this poor performance is the diagonal penalty. Both methods use a non-convex diagonal penalty of the same form as the off-diagonal penalty. As discussed in Section 2.8, setting diagonal penalties in such a way is not optimal. In this case, one effect is that small diagonal entries receive large penalisation and shrinkage, whilst large diagonal entries receive small penalisation and shrinkage. This may explain the poor performance when data is standardised by S.

SCAD and MCP also tend to have poor performance when the sample size n is small relative to the dimension p, as in the case in this gene expression example. One reason for this might be that when the sample size is small, it is less clear from the data which of the entries of Θ are large. Hence the non-convex penalty, designed to reduce bias in the estimation of large entries of Θ, is less effective. Such a phenomenon will be demonstrated again in the applications of Chapter 4.

When data is standardised by S^{-1}, all of GLASSO, SCAD and MCP perform significantly worse than GLASSO with no diagonal penalty. The reason for this may be that the diagonal penalty in these methods uses the same regularisation parameter as the off-diagonal penalties. In order to obtain sparsity in the estimate, one must use a large regularisation parameter. However, this will cause a larger penalisation of the diagonal entries, shrinking them far from their true value and result in decreased predictive performance.

2.10.3 Stock market data

We now perform the same analysis on the stock market data in the R package **huge**, investigated in the graphical model context by Banerjee and Ghosal [2015]. The data contains daily closing stock prices of companies in the S&P 500 index between 1st January 2003 and 1st January 2008. We consider de-trended stock-market log-returns, to study the dependence structure after accounting for the overall mean market behavior. Specifically, let Y_{jt} be the closing price of company j at time t, $\tilde{X}_{jt} = \log\left(\frac{Y_{j,t+1}}{Y_{jt}}\right)$ the log-returns, and $X_{jt} = \tilde{X}_{jt} - \bar{X}_t$ the de-trended returns, where $\bar{X}_t = \sum_{j=1}^{p} \tilde{X}_{jt}$. We randomly selected $p = 30$ companies and, to avoid issues with stock market data exhibiting thicker tails than the assumed Gaussian model, we removed outlying observations more than 5 sample standard deviations away from the mean in any of the p variables. There remained 1,121 observations of which we

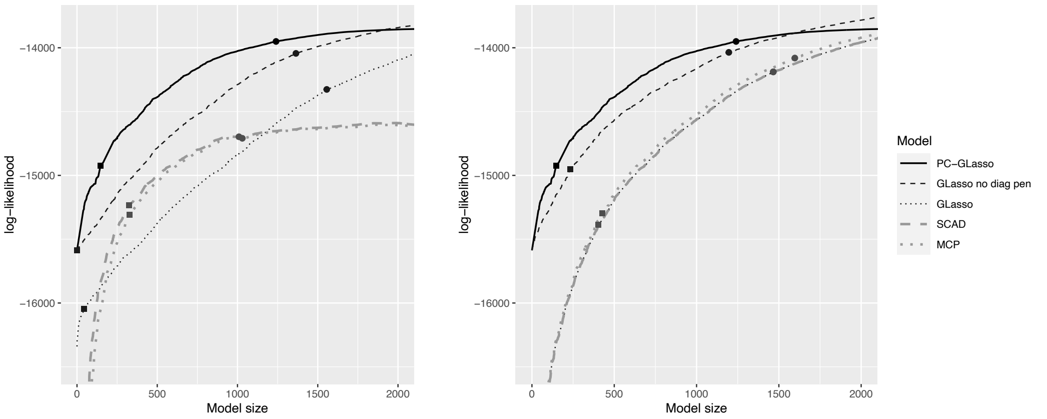


Figure 2.12: Model size vs predictive ability in the gene expression data. Left shows methods on data standardised by S, right shows methods on data standardised by S^{-1}. Estimates selected via BIC and EBIC with $\gamma = 0.5$ are shown by dots and squares respectively.

randomly selected 1,000 for the training and 121 for the test data.

One consideration to make about this stock market data is the appropriateness of the assumed Gaussian model. We have already noted that this data displays thicker tails than the Gaussian model and have removed outlying observations to reduce this affect. An important aspect of methods for Gaussian graphical model is robustness to non-Gaussian settings since this assumption will be violated to some degree in most real data settings. A potentially more troublesome aspect is that the dependencies between stock market prices may not remain constant over time - that is the observations may not be identically distributed. We have attempted to mitigate this by taking test data randomly over time.

The results are shown in Figure 2.13, again on data standardised by S in the left panel and standardised by S^{-1} in the right. This time there is less of a difference between the five methods, and also less of a difference between the two standardisations. In both standardisations PC-GLASSO achieves a slightly higher log-likelihood for all model sizes smaller than 200. The BIC and EBIC estimates are also more sparse than the GLASSO estimates, although with a slight trade off in predictive ability. SCAD and MCP perform better in this example, selecting the simplest model with the BIC estimates with no loss in predictive ability. One reason for this may be the relatively large sample size in this example. However, their predictive ability is worse for larger model sizes.

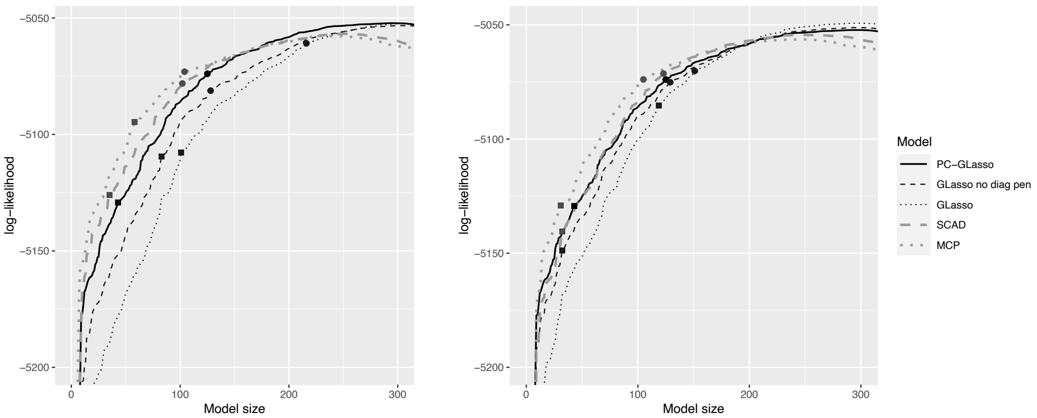


Figure 2.13: Model size vs predictive ability in the stock market data. Left shows methods on data standardised by S, right shows methods on data standardised by S^{-1}. Estimates selected via BIC and EBIC with $\gamma = 0.5$ are shown by dots and squares respectively.

2.11 Discussion

Penalised likelihood methods based on regular penalty functions are a staple of Gaussian graphical model selection and precision matrix estimation. They provide a conceptually easy strategy to obtain sparse estimates of Θ and, particularly in the case of GLASSO, fairly efficient computation, even for moderately large dimensions. However, in this paper we demonstrated that estimates obtained from regular penalties depend on the scale of the variables. This gives a situation where a simple change of units (measuring a distance in miles rather than kilometers) can result in different graphical model selection. Further, we introduced a notion of exchangeability motivating the need for standardising the data when using regular penalties.

Standardising the data is not innocuous. First, even when the original variables follow a Gaussian distribution, that is no longer the case once the variables have been standardised. They then exhibit thicker tails [Kollo and Ruul, 2003]. Second, from a more applied viewpoint, as demonstrated in several of our examples, applying regular penalties to scaled data can adversely affect inference. Through simulation experiments we were able to demonstrate that this effect was particularly detrimental in examples when the true underlying graph has a large range in node degrees, as in the Star graph setting.

To combat these issues we have made two key recommendations in this section which should be applied to all penalised likelihood methods in Gaussian graphical models. First, the penalty should be a function of the partial correlations rather than the off-diagonal entries of the precision matrix. Second, the penalty on the the diagonal entries of the precision matrix should either be logarithmic or not penalised at all. We have shown that all penalised likelihood methods that follow these two suggestions satisfy scale invariance and so, for example, will result in the same inference regardless of the units of measurement or standardisation.

We also investigated one such penalty function, the PC-GLASSO, which sets an L_1 penalty on the partial correlations. This is a direct parallel to the regular GLASSO penalty which sets an L_1 penalty on Θ. First we demonstrated that the PC-GLASSO provides a preferable shrinkage of the partial correlations and diagonal entries in the $p = 2$ case. We then showed through simulated examples that PC-GLASSO can lead to significant improvements in both estimation and model selection over GLASSO in certain settings, and also offers improved estimation in all small sample settings that we explored.

A current limitation of our work lies in the computational efficiency of our coordinate descent algorithm. While the efficiency of this algorithm is reasonable in lower dimensions, the computations become impractical for larger p. However, the conditional convexity of the PC-GLASSO problem opens interesting strategies for future improvements. For example, one may alternately update the estimated diagonal entries $\hat{\theta}$ and partial correlations $\hat{\Delta}$, keeping the other fixed. For fixed $\hat{\theta}$, widely studied computational methods for regular penalties may be employed to maximise the penalised likelihood for Δ. With such a computational method, the speed of PC-GLASSO may be competitive with GLASSO and could also be applied to large scale problems.

Further interesting future work is to investigate the theoretical properties of PC-GLASSO, for example model selection consistency, which holds for GLASSO only under certain nontrivial conditions [Ravikumar et al., 2009]. The wider set of PC-separable penalties also warrant further exploration, most obviously PC-separable versions of the SCAD and MCP penalties. Beyond the Gaussian case, penalisation of partial correlations also seems natural for partial correlation graphs in elliptical and transelliptical distributions, see Rossell and Zwiernik [2020].

$n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
M1	**1.42 (0.35)**	**1.69 (0.58)**	**0.978 (0.043)**	0.999 (0.008)	0.995 (0.010)
M2	1.81 (0.58)	2.05 (0.78)	0.966 (0.061)	0.998 (0.009)	0.992 (0.015)
M3	2.68 (0.73)	3.55 (1.19)	0.231 (0.063)	0.903 (0.066)	0.477 (0.075)
M4	1.93 (0.29)	2.57 (0.56)	0.569 (0.141)	0.988 (0.027)	0.810 (0.107)
M5	2.59 (0.36)	3.36 (1.42)	0.270 (0.044)	0.866 (0.105)	0.578 (0.057)
M6	1.49 (0.28)	2.01 (0.69)	0.789 (0.127)	0.996 (0.016)	0.934 (0.054)

$n = 100$	FNorm	KL	MCC	Sensitivity	Specificity
M1	**0.70 (0.11)**	**0.46 (0.12)**	**0.993 (0.017)**	1 (0)	0.999 (0.004)
M2	0.73 (0.13)	0.48 (0.13)	**0.993 (0.018)**	1 (0)	0.998 (0.004)
M3	1.73 (0.08)	1.33 (0.13)	0.264 (0.021)	0.996 (0.014)	0.433 (0.041)
M4	1.30 (0.14)	1.07 (0.24)	0.679 (0.103)	1 (0)	0.887 (0.056)
M5	1.66 (0.09)	1.20 (0.13)	0.304 (0.019)	0.996 (0.014)	0.508 (0.031)
M6	0.84 (0.12)	0.62 (0.16)	0.907 (0.060)	1 (0)	0.978 (0.016)

Table 2.1: Star results

$n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
M1	**1.85 (0.29)**	**2.83 (0.74)**	0.696 (0.081)	0.988 (0.043)	0.917 (0.034)
M2	2.18 (0.49)	3.25 (0.92)	**0.714 (0.089)**	0.986 (0.045)	0.924 (0.034)
M3	2.51 (0.28)	3.71 (0.70)	0.371 (0.066)	0.999 (0.009)	0.644 (0.095)
M4	2.33 (0.26)	3.52 (0.65)	0.438 (0.081)	0.998 (0.012)	0.726 (0.089)
M5	2.26 (0.21)	3.11 (0.64)	0.469 (0.071)	0.998 (0.012)	0.763 (0.066)
M6	2.00 (0.23)	2.95 (0.66)	0.560 (0.077)	0.996 (0.018)	0.839 (0.054)

$n = 100$	FNorm	KL	MCC	Sensitivity	Specificity
M1	0.91 (0.15)	**0.70 (0.20)**	0.858 (0.069)	1 (0)	0.969 (0.019)
M2	**0.89 (0.15)**	**0.70 (0.21)**	**0.878 (0.063)**	1 (0)	0.974 (0.016)
M3	1.84 (0.19)	1.37 (0.20)	0.371 (0.038)	1 (0)	0.650 (0.054)
M4	1.65 (0.23)	1.26 (0.27)	0.449 (0.058)	1 (0)	0.743 (0.057)
M5	1.63 (0.19)	1.11 (0.22)	0.483 (0.054)	1 (0)	0.778 (0.048)
M6	1.40 (0.20)	1.01 (0.24)	0.631 (0.074)	1 (0)	0.882 (0.038)

Table 2.2: Hub results

$n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
M1	3.64 (0.31)	**5.26 (0.62)**	0.283 (0.093)	0.301 (0.194)	0.922 (0.077)
M2	**3.42 (0.34)**	5.84 (0.97)	**0.296 (0.080)**	0.371 (0.168)	0.891 (0.080)
M3	4.27 (0.17)	6.63 (0.71)	0.258 (0.113)	0.162 (0.135)	0.978 (0.041)
M4	4.08 (0.28)	6.05 (0.75)	0.268 (0.083)	0.297 (0.149)	0.913 (0.084)
M5	3.83 (0.15)	5.77 (0.55)	0.219 (0.128)	0.126 (0.140)	0.984 (0.032)
M6	3.77 (0.16)	5.66 (0.56)	0.212 (0.104)	0.148 (0.137)	0.970 (0.042)

$n = 100$	FNorm	KL	MCC	Sensitivity	Specificity
M1	2.30 (0.33)	**2.00 (0.38)**	0.530 (0.052)	0.855 (0.094)	0.774 (0.069)
M2	**2.18 (0.34)**	2.03 (0.41)	**0.537 (0.055)**	0.851 (0.100)	0.783 (0.068)
M3	2.70 (0.45)	2.10 (0.52)	0.462 (0.062)	0.903 (0.090)	0.663 (0.112)
M4	2.72 (0.44)	2.15 (0.52)	0.468 (0.061)	0.893 (0.090)	0.678 (0.112)
M5	2.72 (0.36)	2.22 (0.49)	0.520 (0.057)	0.818 (0.122)	0.785 (0.092)
M6	2.68 (0.34)	2.18 (0.45)	0.526 (0.053)	0.835 (0.098)	0.781 (0.079)

Table 2.3: AR2 results

$n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
M1	**2.30 (0.25)**	**3.07 (0.51)**	0.336 (0.091)	0.310 (0.153)	0.951 (0.041)
M2	2.40 (0.41)	3.40 (0.67)	0.337 (0.080)	0.339 (0.143)	0.939 (0.044)
M3	2.84 (0.19)	4.32 (0.63)	**0.355 (0.085)**	0.264 (0.136)	0.969 (0.048)
M4	2.66 (0.22)	3.82 (0.61)	0.312 (0.079)	0.367 (0.134)	0.915 (0.063)
M5	2.38 (0.18)	3.49 (0.61)	0.311 (0.118)	0.205 (0.158)	0.978 (0.040)
M6	2.31 (0.20)	3.33 (0.62)	0.278 (0.105)	0.224 (0.148)	0.964 (0.038)

$n = 100$	FNorm	KL	MCC	Sensitivity	Specificity
M1	1.43 (0.16)	**1.23 (0.25)**	0.572 (0.059)	0.614 (0.110)	0.941 (0.029)
M2	**1.36 (0.14)**	**1.23 (0.25)**	0.571 (0.061)	0.612 (0.112)	0.941 (0.030)
M3	1.93 (0.22)	1.64 (0.37)	0.526 (0.070)	0.724 (0.102)	0.871 (0.065)
M4	1.86 (0.20)	1.54 (0.31)	0.514 (0.066)	0.706 (0.100)	0.876 (0.049)
M5	1.62 (0.15)	1.37 (0.27)	**0.581 (0.061)**	0.641 (0.102)	0.941 (0.030)
M6	1.57 (0.16)	1.32 (0.28)	0.559 (0.062)	0.613 (0.115)	0.936 (0.031)

Table 2.4: Random graph results

Chapter 3

Bayesian partial correlation graphical LASSO

There is a well known equivalence between penalised likelihood estimates and the maximum a-posteriori (MAP) estimate under certain prior distributions. Let $\pi(\Theta)$ be a prior density on the precision matrix Θ and S be a sample covariance matrix associated to observations of a Gaussian graphical model, as in Chapter 2. Then the posterior density can be written as

$$\pi(\Theta \mid S) \propto L(\Theta \mid S)\pi(\Theta),$$

where L is the Gaussian likelihood function for the precision matrix Θ. The MAP estimate is the value of Θ which maximises the posterior density, or equivalently its logarithm

$$\log\left(L(\Theta \mid S)\pi(\Theta)\right) = l(\Theta \mid S) + \log\left(\pi(\Theta)\right),$$

where l is the log-likelihood. Notice that this is of the same form as the penalised likelihood function

$$l(\Theta \mid S) - Pen(\Theta).$$

Hence the MAP estimate under the prior π is equal to the penalised likelihood estimate under the penalty function $Pen(\Theta) = -\log\left(\pi(\Theta)\right).$

Note that prior densities are restricted to integrate to 1 (or have unbounded integral if the prior is improper) and must be equal to 0 on matrices that are not positive definite and symmetric. So more generally the penalised likelihood estimate under the penalty function $Pen(\Theta)$ is equal to the MAP estimate under the prior density

$$\pi(\Theta) \propto \exp\left(-Pen(\Theta)\right)\mathbb{I}(\Theta \in \mathcal{S})$$

where $\mathcal{S}$ is the set of symmetric, positive definite matrices.

This equivalence is useful in a number of ways. Primarily, it provides a Bayesian framework for penalised likelihood methods allowing full posterior analysis. Unlike the penalised likelihood, which only provides a point estimate for Θ, the posterior distribution quantifies posterior uncertainty in the values of Θ. This was utilised by Wang [2012] and Khondker et al. [2013] for the Bayesian parallel of the GLASSO penalty function. However, posterior analysis is often more challenging due to the form of the posterior density. In particular, for many prior distributions, the posterior marginals do not have a closed form. Posterior inference therefore often requires a sampling scheme, which can be computationally expensive for large dimension size p. Furthermore, if the prior distribution is continuous, then the event $\{\theta_{ij} = 0\}$ will have zero posterior probability. The posterior distribution will therefore have no direct way to assign edge inclusion probabilities, or to conduct graphical model selection. One may compare Bayes factors associated to different graphical models [Consonni et al., 2018], however for large p an exhaustive search is not computationally feasible.

Beyond posterior inference, the Bayesian parallel to penalised likelihoods can be useful in other ways. For example, the prior distribution gives an alternative interpretation to the penalty function which can provide additional insight into the dynamics of the penalised likelihood. The prior distribution associated to GLASSO was explored by Wang [2012] showing that, for example, the marginal densities of the partial correlations become more concentrated around 0 for larger dimension size p due to the positive definiteness truncation in the prior.

The Bayesian interpretation can also provide inspiration for new penalty functions. When, for computational reasons, full posterior analysis is avoided in favour of simply finding a MAP estimate, the Bayesian method can be equivalently seen as a penalised likelihood. However, the prior distribution provides a different framework for interpretation that can lead to new ideas for penalty functions. For example, Banerjee and Ghosal [2015] and Gan et al. [2018] both proposed methods which involve finding the MAP estimate associated to a spike and slab prior distribution. Marlin et al. [2009] and Marlin and Murphy [2009] also utilised the prior interpretation to perform inference on block structured precision matrices.

In this chapter we explore prior distributions associated to regular and PC-separable penalty functions, paying particular attention to the GLASSO and PC-GLASSO. We do not provide any concrete means of posterior inference in the case of PC-GLASSO, instead focusing on the interpretative aspect of the prior distribution.

The remainder of the chapter is organised as follows. In Section 3.1 we introduce the class of prior distributions related to regular penalty functions with the GLASSO prior as a particular case. In Section 3.2 we introduce the class of prior distributions related to PC-separable penalty functions with the PC-GLASSO prior as a specific case. In Section 3.3 we show that certain PC-separable priors satisfy an extended Bayesian form of scale invariance. In Section 3.4 we review current research into the GLASSO prior and in Section 3.5 compare to the PC-GLASSO prior.

3.1 Separable prior distributions

Consider the same Gaussian graphical model set up as in Chapter 2 described in Section 2.1. We define a class of prior distributions on Θ, called *seperable priors* in which the elements of Θ are independent, up to positive definiteness and symmetry. We then define a subclass of priors, called *regular priors* which imposes symmetry and differentiability constraints.

Definition 7. A prior distribution with (possibly improper) density $\pi(\Theta)$ is *separable* if

$$\pi(\Theta) \propto \prod_{i \leq j} \pi_{ij}(\theta_{ij}) \mathbb{I}(\Theta \in \mathcal{S}) \tag{3.1}$$

where $\pi_{ii} : (0, \infty) \to [0, \infty)$ and $\pi_{ij} : \mathbb{R} \to [0, \infty)$ are non-increasing in θ_{ii} and $|\theta_{ij}|$ respectively for all i and $i < j$.

A separable prior distribution is *regular* if $\pi_{ii} = \pi_{jj}$ for all (i, j) and for all $i < j$, π_{ij} does not depend on (i, j), is symmetric about 0 and differentiable away from 0.

Although not strictly necessary, it is useful to require that that π_{ij} be (possibly improper) density functions. In this way the π_{ij} represent the marginal prior densities before truncation onto the space of positive definite matrices and the proportionality in (3.1) comes only from this truncation. From now on we will assume that such π_{ij} are density functions, as well as in the later Definition 8.

Regular priors are often a reasonable choice for representing prior beliefs in a simple manner. They have identical marginals on each of the θ_{ij} and each of the θ_{ii} which represents a symmetry in the prior knowledge about each edge in the graph - that is, the a priori knowledge of each edge in the graph is identical for all edges. The separable form of the prior also allows a simple framework to set prior beliefs on the magnitude of the θ_{ij}.

Note that although a separable prior density is written as the product of functions only depending on single entries of Θ, these do not correspond to the marginal densities and the entries of Θ are not independent under this prior due to the truncation onto the space of symmetric, positive definite matrices.

Proposition 6. *If the penalty function Pen is separable (regular), then the prior density $\pi(\Theta) \propto \exp\left(-Pen(\Theta)\right)\mathbb{I}(\Theta \in \mathcal{S})$ is separable (regular).*

If the prior density $\pi(\Theta) \propto \prod_{i\leq j} \pi_{ij}(\theta_{ij})\mathbb{I}(\Theta \in \mathcal{S})$ is separable (regular), then the penalty function $Pen(\Theta) = -\log\left(\prod_{i\leq j} \pi_{ij}(\theta_{ij})\right)$ is separable (regular).

Proof. Let Pen be a separable penalty function and $\pi(\Theta) \propto \exp\left(-Pen(\Theta)\right)\mathbb{I}(\Theta \in \mathcal{S})$. Then

$$\begin{aligned}\pi(\Theta) &\propto \exp\left(-\sum_{i\leq j} pen_{ij}(\theta_{ij})\right)\mathbb{I}(\Theta \in \mathcal{S})\\ &= \prod_{i\leq j}\exp\left(-pen_{ij}(\theta_{ij})\right)\mathbb{I}(\Theta \in \mathcal{S})\\ &= \prod_{i\leq j}\pi_{ij}(\theta_{ij})\mathbb{I}(\Theta \in \mathcal{S})\end{aligned}$$

where $\pi_{ij}(\theta_{ij}) = \exp\left(-pen_{ij}(\theta_{ij})\right)$. Hence $\pi(\Theta)$ is of the form (3.1), π_{ij} has range $[0,\infty)$ for all (i,j) and, because $pen_{ij}(\theta_{ij})$ is non-decreasing in $|\theta_{ij}|$, $\pi_{ij}(\theta_{ij})$ is non-increasing in $|\theta_{ij}|$. Therefore π is separable.

If Pen is also regular, it easily follows that $\pi_{ii} = \pi_{jj}$ for all (i,j) and that π_{ij} does not depend on (i,j) for all $i < j$, is symmetric about 0 and differentiable away from 0. Hence π is also regular.

Now suppose that $\pi(\Theta) \propto \prod_{i\leq j} \pi_{ij}(\theta_{ij})\mathbb{I}(\Theta \in \mathcal{S})$ is a separable (regular) prior density and let $Pen(\Theta) = -\log\left(\prod_{i\leq j} \pi_{ij}(\theta_{ij})\right)$. Then

$$\begin{aligned}Pen(\Theta) &= -\sum_{i\leq j}\log\left(\pi_{ij}(\theta_{ij})\right)\\ &= \sum_{i\leq j} pen_{ij}(\theta_{ij})\end{aligned}$$

where $pen_{ij}(\theta_{ij}) = -\log\left(\pi_{ij}(\theta_{ij})\right)$. It easily follows that Pen is a separable (regular) penalty function.

□

Proposition 6 shows that every regular penalty function corresponds to a regular prior distribution and vice versa. Note that any two regular penalties that are proportional to one another correspond to the same regular prior. The statement of Proposition 6 was written in this form, rather than an if and only if statement, because penalty functions are not restricted to positive definite matrices. The if and only if statement can be obtained by simply making this restriction in the definition of the penalty function.

In this chapter we will primarily focus on the regular prior distribution corresponding to the GLASSO penalty function. Recall that the GLASSO penalty has diagonal penalty $pen_{ii}(\theta_{ii}) = \frac{n}{2}\rho\theta_{ii}$ and off-diagonal penalty $pen_{ij}(\theta_{ij}) = n\rho|\theta_{ij}|$. The corresponding prior, which we refer to as the GLASSO prior, with parameter $\lambda = n\rho$ therefore has density

$$\begin{aligned}\pi_{\mathrm{G}}(\Theta|\lambda) &\propto \exp\left(-\frac{1}{2}\lambda\sum_i \theta_{ii} - \lambda\sum_{i<j}|\theta_{ij}|\right)\mathbb{I}(\Theta \in \mathcal{S})\\ &= \prod_i \exp\left(-\frac{1}{2}\lambda\theta_{ii}\right)\prod_{i<j}\exp\left(-\lambda|\theta_{ij}|\right)\mathbb{I}(\Theta \in \mathcal{S})\\ &\propto \prod_i \mathrm{Exp}(\theta_{ii};\lambda/2)\prod_{i<j}\mathrm{Laplace}(\theta_{ij};0,\lambda^{-1})\mathbb{I}(\Theta \in \mathcal{S})\end{aligned}$$

where $\mathrm{Exp}(\theta_{ii};\lambda/2)$ denotes the density of an exponential distribution with rate parameter $\lambda/2$ and $\mathrm{Laplace}(\theta_{ij};0,\lambda^{-1})$ denotes the density of a Laplace distribution with location 0 and scale λ^{-1}.

Note that while the GLASSO prior density can be written as the product of exponential and Laplace densities, the positive definite truncation means that the marginal densities do not have these forms. This will be explored further in Sections 3.4 and 3.5.

An additional interesting regular prior distribution is used in the graphical horseshoe introduced by Li et al. [2019]. In the graphical horseshoe, $\pi_{ii}(\theta_{ii}) \propto 1$ is an improper uniform and $\pi_{ij}(\theta_{ij}) = \mathrm{N}(\theta_{ij};0,\lambda_{ij}^2\tau^2)\mathrm{C}^+(\lambda_{ij};0,1)$ where $\mathrm{C}^+(x;0,1) \propto (1+x^2)^{-1}$ is the half-Cauchy density. Conditional on the hyperparameter τ, this results in a regular prior distribution. In the graphical horseshoe, τ is also given a half-Cauchy hyper prior. The combination of the Normal distribution with variance parameter determined by two half-Cauchy distributions results in a distribution which is highly peaked close to 0 but with slowly decaying tails.

Since in a regular prior distribution the prior is continuous, direct graphical model selection is not possible because under the posterior distribution $\theta_{ij} \neq 0$ almost surely. One option is to only consider the MAP estimation which, for certain choices of regular priors, can lead to sparse estimation. In the Bayesian GLASSO and graphical horseshoe, instead the posterior mean estimate is considered which is not sparse. In the graphical horseshoe model selection is conducted by considering posterior credible intervals and checking if these contain 0. Model selection in the Bayesian GLASSO will be discussed in Section 3.4.

3.2 Partial correlation separable prior distributions

We now propose an alternative class of prior distributions which are instead separable in the partial correlations.

Definition 8. A prior distribution with (possibly improper) density $\pi(\theta, \Delta)$ is (symmetric) PC-separable if

$$\pi(\theta, \Delta) \propto \prod_{i} \pi_{ii}(\theta_{ii}) \prod_{i<j} \pi_{ij}(\Delta_{ij}) \mathbb{I}(\Delta \in \mathcal{S}_1), \tag{3.2}$$

where $\pi_{ii} : (0, \infty) \to [0, \infty)$ and $\pi_{ij} : [-1, 1] \to [0, \infty)$ are non-increasing in θ_{ii} and $|\Delta_{ij}|$ respectively for all i and $i < j$ and $\mathcal{S}_1$ denotes the set of symmetric, positive definite matrices with unit diagonal.

A PC-separable prior distribution is *symmetric* if $\pi_{ii} = \pi_{jj}$ for all (i, j) and for all $i < j$, pen_{ij} does not depend on (i, j) and is symmetric about 0.

Symmetric PC-separable priors benefit from the same simple interpretation as regular priors - under a symmetric PC-separable prior the a priori knowledge of each edge is identical and the separable form allows simple setting of prior beliefs on the magnitude of the partial correlations.

Note that in a PC-separable prior, the truncation only restricts Δ to be positive definite. However, this ensures that Θ is positive definite because $\Theta = \theta^{1/2}\Delta\theta^{1/2}$ and θ is a diagonal matrix with strictly positive entries. Therefore Θ is positive definite if and only if Δ is positive definite.

Although a PC-separable prior density is written as the product of functions of individual entries of Δ, these do not correspond to the marginal densities and the partial correlations are not independent under the prior due to the truncation onto the space of symmetric, positive definite matrices with unit diagonal. However, there is no truncation on the diagonal entries, and so the marginal density of θ_{ii} is given by π_{ii} and the diagonal entries are independent of one another and of the partial correlations under a PC-separable prior.

Proposition 7. *If the penalty function Pen is (symmetric) PC-separable, then the prior density* $\pi(\theta, \Delta) \propto \exp\left(-Pen(\theta, \Delta)\right) \mathbb{I}(\Delta \in \mathcal{S}_1)$ *is (symmetric) PC-separable.*

If the prior density $\pi(\theta, \Delta) \propto \prod_i \pi_{ii}(\theta_{ii}) \prod_{i<j} \pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta \in \mathcal{S}_1)$ is (symmetric) PC-separable, then the penalty function $Pen(\theta, \Delta) = -\log\left(\prod_i \pi_{ii}(\theta_{ii}) \prod_{i<j} \pi_{ij}(\Delta_{ij})\right)$ is (symmetric) PC-separable.

Proof. The result follows easily (and in a similar manner to that of Proposition 6) by noting the following.

If Pen is a PC-separable penalty function, then $\pi(\theta, \Delta) \propto \exp\left(-Pen(\theta, \Delta)\right)\mathbb{I}(\Delta \in \mathcal{S}_1)$ can be written as

$$\begin{aligned}\pi(\theta, \Delta) &\propto \exp\left(-\sum_i pen_{ii}(\theta_{ii}) - \sum_{i<j} pen_{ij}(\Delta_{ij})\right)\mathbb{I}(\Delta \in \mathcal{S}_1)\\ &= \prod_i \exp\left(-pen_{ii}(\theta_{ii})\right) \prod_{i<j} \exp\left(-pen_{ij}(\Delta_{ij})\right)\mathbb{I}(\Delta \in \mathcal{S}_1)\\ &= \prod_i \pi_{ii}(\theta_{ii}) \prod_{i<j} \pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta \in \mathcal{S}_1)\end{aligned}$$

where $\pi_{ii}(\theta_{ii}) = \exp\left(-pen_{ii}(\theta_{ii})\right)$ and $\pi_{ij}(\Delta_{ij}) = \exp\left(-pen_{ij}(\Delta_{ij})\right)$.

Conversely, if $\pi(\theta, \Delta) \propto \prod_i \pi_{ii}(\theta_{ii}) \prod_{i<j} \pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta \in \mathcal{S}_1)$ is a PC-separable prior density, then

$$\begin{aligned}Pen(\theta, \Delta) &= -\log\left(\prod_i \pi_{ii}(\theta_{ii}) \prod_{i<j} \pi_{ij}(\Delta_{ij})\right)\\ &= -\sum_i \log(\pi_{ii}(\theta_{ii})) - \sum_{i<j} \log(\pi_{ij}(\Delta_{ij}))\\ &= \sum_i pen_{ii}(\theta_{ii}) + \sum_{i<j} pen_{ij}(\Delta_{ij})\end{aligned}$$

where $pen_{ii}(\theta_{ii}) = -\log(\pi_{ii}(\theta_{ii}))$ and $pen_{ij}(\Delta_{ij}) = -\log(\pi_{ij}(\Delta_{ij}))$.

□

In this chapter the PC-seperable prior we will primarily focus on is that relating to the PC-GLASSO penalty, however a PC-separable prior was also proposed by Wong and Carter [2003]. Recall that the PC-GLASSO penalty has diagonal penalty $pen_{ii}(\theta_{ii}) = 2\log(\theta_{ii})$ and partial correlation penalty $pen_{ij}(\Delta_{ij}) = n\rho|\Delta_{ij}|$. The corresponding prior, which we refer to as the PC-GLASSO prior, with parameter

$\lambda = n\rho$ therefore has density

$$\begin{aligned}
\pi_{\mathrm{PC}}(\theta, \Delta|\lambda) &\propto \exp\left(\sum_i \log(\theta_{ii}^{-2}) - \lambda \sum_{i<j} |\Delta_{ij}|\right) \mathbb{I}(\Delta \in \mathcal{S}_1) \\
&= \prod_i \theta_{ii}^{-2} \prod_{i<j} \exp\left(-\lambda|\Delta_{ij}|\right) \mathbb{I}(\Delta \in \mathcal{S}_1) \\
&\propto \prod_i \theta_{ii}^{-2} \prod_{i<j} \mathrm{Laplace}(\Delta_{ij}; 0, \lambda^{-1}) \mathbb{I}(\Delta \in \mathcal{S}_1)
\end{aligned}$$

Although the prior density is written in terms of a Laplace density on the partial correlations, the truncation $\mathbb{I}(\Delta \in \mathcal{S}_1)$ means that the marginal densities are not Laplace. In particular, the truncation ensures that the partial correlations are in $(-1, 1)$. However, the marginal densities on the diagonal entries are given by $\pi_{\mathrm{PC}}(\theta_{ii}) \propto \theta_{ii}^{-2}$. It can easily be shown that this implies an improper uniform distribution on the partial variances θ_{ii}^{-1}, which gives a nice interpretation of this prior.

The PC-GLASSO prior is improper due to the improper marginals on the diagonal entries. However, a regular version can be obtained by restricting $\theta_{ii} > \epsilon$ for any $\epsilon > 0$. It is unclear whether the posterior distribution corresponding to the PC-GLASSO prior is in general proper, however it is for a certain subset of sample covariance matrices S (see below). In this chapter we focus on the interpretation of the PC-GLASSO prior to better understand the corresponding penalised likelihood, as opposed to conducting inference based on the posterior. As such, having a proper posterior is not vital to the content of this chapter.

Proposition 8. *For any sample size $n \geq 4$ and sample covariance matrix S satisfying $S_{ii} > \sum_{j \neq i} |S_{ij}|$ for all i, the posterior distribution associated to the PC-GLASSO prior is proper.*

Proof. The posterior density is written as

$$\begin{aligned}
\pi_{\mathrm{PC}}(\theta, \Delta|\lambda, S) &\propto \pi_{\mathrm{PC}}(\theta, \Delta|\lambda) L(\theta, \Delta|S) \\
&\propto \prod_i \theta_{ii}^{-2} \prod_{i<j} \exp\left(-\lambda|\Delta_{ij}|\right) \\
&\qquad \times \prod_i \theta_{ii}^{n/2} \det(\Delta)^{n/2} \exp\left(-\frac{n}{2}\sum_{i,j} S_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}\right) \mathbb{I}(\Delta \in \mathcal{S}_1) \\
&= \det(\Delta)^{n/2} \prod_i \left(\theta_{ii}^{\frac{n}{2}-2} \exp\left(-\frac{n}{2} S_{ii}\theta_{ii}\right)\right) \\
&\qquad \times \prod_{i<j} \exp\left(-\lambda|\Delta_{ij}| - nS_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}\right) \mathbb{I}(\Delta \in \mathcal{S}_1)
\end{aligned}$$

We must show that this function integrates to a finite value. First note that $\det(\Delta)^{n/2} \leq 1$ since $\log(\det(\Delta)) \leq \mathrm{tr}(\Delta - I)$, $\exp(-\lambda|\Delta_{ij}|) \leq 1$ since $\lambda > 0$, and $\exp(-nS_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}) < \exp(n|S_{ij}|\sqrt{\theta_{ii}\theta_{jj}}) \leq \exp(\frac{n}{2}|S_{ij}|(\theta_{ii} + \theta_{jj}))$ because $-1 < \Delta_{ij} < 1$ and $\sqrt{\theta_{ii}\theta_{jj}} \leq \frac{1}{2}(\theta_{ii} + \theta_{jj})$. Hence

$$\begin{aligned}
&\int_{\mathcal{S}_1}\int_{\mathbb{R}_+^p} \det(\Delta)^{n/2} \prod_i \left(\theta_{ii}^{\frac{n}{2}-2} \exp\left(-\frac{n}{2}S_{ii}\theta_{ii}\right)\right) \prod_{i<j} \exp\left(-\lambda|\Delta_{ij}| - nS_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}\right) d\theta \, d\Delta \\
&\quad\leq \int_{\mathcal{S}_1}\int_{\mathbb{R}_+^p} \prod_i \left(\theta_{ii}^{\frac{n}{2}-2} \exp\left(-\frac{n}{2}S_{ii}\theta_{ii}\right)\right) \prod_{i<j} \exp\left(\frac{n}{2}|S_{ij}|(\theta_{ii} + \theta_{jj})\right) d\theta \, d\Delta \\
&\quad= \int_{\mathcal{S}_1}\int_{\mathbb{R}_+^p} \prod_i \left(\theta_{ii}^{\frac{n}{2}-2} \exp\left(-\frac{n}{2}S_{ii}\theta_{ii}\right)\right) \prod_i \exp\left(\frac{n}{2}\sum_{j\neq i}|S_{ij}|\theta_{ii}\right) d\theta \, d\Delta \\
&\quad= \int_{\mathcal{S}_1}\int_{\mathbb{R}_+^p} \prod_i \theta_{ii}^{\frac{n}{2}-2} \exp\left(-\frac{n}{2}\left(S_{ii} - \sum_{j\neq i}|S_{ij}|\right)\theta_{ii}\right) d\theta \, d\Delta \\
&\quad= \int_{\mathcal{S}_1}\left(\prod_i \int_{\mathbb{R}_+} \theta_{ii}^{\frac{n}{2}-2} \exp\left(-\frac{n}{2}\left(S_{ii} - \sum_{j\neq i}|S_{ij}|\right)\theta_{ii}\right) d\theta_{ii}\right) d\Delta
\end{aligned}$$

Since the function being integrated does not depend on Δ and $\mathcal{S}_1$ is a bounded set, if the inner integral is finite then the whole integral is also finite. Also, whenever $\frac{n}{2} - 2 \geq 0$ and $S_{ii} > \sum_{j\neq i}|S_{ij}|$, then $\int_{\mathbb{R}_+} \theta_{ii}^{\frac{n}{2}-2} \exp\left(-\frac{n}{2}\left(S_{ii} - \sum_{j\neq i}|S_{ij}|\right)\theta_{ii}\right) d\theta_{ii}$ is finite. It follows that the whole integral is finite.

□

We have shown that under a certain condition on the sample covariance matrix, the PC-GLASSO posterior distribution is guaranteed to be proper. However, this does not imply that the posterior is improper when this condition does not hold. It may in fact be the case that posterior is proper for any sample covariance as long as $n \geq 4$. To see why this may be the case, consider why the PC-GLASSO prior is improper - the θ_{ii}^{-2} term tends to infinity as θ_{ii} goes to 0. However, in the posterior this is not the case because of the $\theta_{ii}^{n/2}$ term in the likelihood. It may be possible to prove that this is the case, however it is complicated by the interaction term $\sqrt{\theta_{ii}\theta_{jj}}$, the determinant term and the fact that integration is over the space of positive definite matrices.

3.3 Scale invariance

The scale invariance results of Section 2.5 directly apply to the MAP estimate under regular and PC-separable prior distributions. In particular, the MAP estimate under regular prior distributions is not scale invariant. However, the MAP estimate under symmetric PC-separable priors with $\pi_{ii}(\theta_{ii}) \propto \theta_{ii}^{-c}$, $c \geq 0$ is scale invariant. That is, if $\tilde{\Theta} = \hat{\Theta}(DSD)$ is the posterior mode under the scaled sample covariance, then the mode under the original sample covariance is $\hat{\Theta}(S) = D\tilde{\Theta}D$. Hence, the maxima of the two posterior densities are $\pi\left(\tilde{\Theta} \mid DSD\right)$ and $\pi\left(D\tilde{\Theta}D \mid S\right)$.

In fact a stronger property holds for the entire posterior distribution, that PC-separable priors lead to scale-invariant posterior inference, as defined below.

Definition 9. Let $\pi(\Theta)$ be a prior density, S a sample covariance and D a diagonal matrix with non-zero diagonal. Let the posterior density associated to S be $\pi(\Theta \mid S) \propto L(\Theta \mid S)\pi(\Theta)$, and that associated to DSD be $\pi(\Theta \mid DSD) \propto L(\Theta \mid DSD)\pi(\Theta)$ where L is the Gaussian likelihood function.

The prior $\pi(\Theta)$ leads to *scale-invariant posterior inference* if for any (S, D)

$$\mathbb{P}_\pi\left(\Theta \in A \mid DSD\right) = \mathbb{P}_\pi\left(\Theta \in A_D \mid S\right) \tag{3.3}$$

for all measurable sets A where $A_D = \{\Theta : D^{-1}\Theta D^{-1} \in A\}$.

In particular, (3.3) implies that the two posterior distributions on the partial correlations Δ are equal up to appropriate sign changes i.e. when D has all positive entries, $\pi(\Delta \mid S) = \pi(\Delta \mid DSD)$ (since Δ associated to Θ is equal to that associated to $D\Theta D$).

Proposition 9. *Any symmetric PC-separable prior distribution with $\pi_{ii}(\theta_{ii}) \propto \theta_{ii}^{-c}$ for some constant $c \geq 0$ leads to scale-invariant posterior inference, provided the posterior distributions exist.*

Proof. Let π be a prior density as given in Proposition 9, S be some sample covariance and D some diagonal matrix with non-zero entries. Writing $L(\Theta \mid S)$ as the likelihood function, $\Theta = \theta^{1/2}\Delta\theta^{1/2}$ and treating D as a constant, the posteriors given S and DSD are

$$\begin{aligned}
\pi(D\Theta D \mid S) &\propto L(D\Theta D \mid S)\pi(D\Theta D) \\
&\propto \det(\Delta)^{n/2} \prod_i (d_i^2\theta_{ii})^{\frac{n}{2}} \exp\left(-\frac{n}{2}\sum_{i,j} S_{ij}\sqrt{d_i^2\theta_{ii}d_j^2\theta_{jj}}\Delta_{ij}\right) \\
&\quad \times \prod_i (d_i^2\theta_{ii})^{-c} \prod_{ij} \pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta \in \mathcal{S}_1) \\
&= \det(\Delta)^{n/2} \prod_i (d_i^2\theta_{ii})^{\frac{n}{2}-c} \exp\left(-\frac{n}{2}\sum_{i,j} d_i d_j S_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}\right) \\
&\quad \times \prod_{ij} \pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta \in \mathcal{S}_1) \qquad (3.4)
\end{aligned}$$

$$\begin{aligned}
\pi(\Theta \mid DSD) &\propto L(\Theta \mid DSD)\pi(\Theta) \\
&\propto \det(\Delta)^{n/2} \prod_i \theta_{ii}^{\frac{n}{2}} \exp\left(-\frac{n}{2}\sum_{i,j} d_i d_j S_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}\right) \\
&\quad \times \prod_i (\theta_{ii})^{-c} \prod_{ij} \pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta \in \mathcal{S}_1) \\
&= \det(\Delta)^{n/2} \prod_i \theta_{ii}^{\frac{n}{2}-c} \exp\left(-\frac{n}{2}\sum_{i,j} d_i d_j S_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}\right) \\
&\quad \times \prod_{ij} \pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta \in \mathcal{S}_1) \qquad (3.5)
\end{aligned}$$

For any measurable set A and $A_D = \{\Theta : D^{-1}\Theta D^{-1} \in A\}$ the probabilities in Definition 9 can be written as

$$\begin{aligned}
\mathbb{P}_\pi\left(\Theta \in A \mid DSD\right) &= \int_A \pi\left(\Theta \mid DSD\right)\, d\Theta \\
&= \frac{\int_A L(\Theta \mid DSD)\pi(\Theta)\, d\Theta}{\int_{\mathcal{S}} L(\Theta \mid DSD)\pi(\Theta)\, d\Theta}
\end{aligned}$$

and, noting that $\Theta \in A \iff D\Theta D \in A_D$,

$$\begin{aligned}\mathbb{P}_\pi\left(\Theta \in A_D \mid S\right) &= \int_{A_D} \pi\left(\Theta \mid S\right)\, d\Theta \\ &= \int_{A} \pi\left(D\Theta D \mid S\right)\, d\Theta \\ &= \frac{\int_A L(D\Theta D \mid S)\pi(D\Theta D)\, d\Theta}{\int_{\mathcal{S}} L(D\Theta D \mid S)\pi(D\Theta D)\, d\Theta}\end{aligned}$$

The result follows by noting that expression (3.4) can be obtained by multiplying (3.5) by the constant $\prod_i (d_i^2)^{\frac{n}{2}-c}$.

□

We note that the proof of Proposition 9 does not depend on $\pi_{ij}(\Delta_{ij})$ being non-increasing or π_{ij} being the same for all $i \neq j$. Hence the result extends to any prior of the form (3.2) for which $\pi_{ii}(\theta_{ii}) \propto \theta_{ii}^{-c}$ for all i and $\pi_{ij}(\Delta_{ij})$ is symmetric for all $i \neq j$. The symmetry condition for π_{ij} is required for negative scalar multiplications - i.e. when D includes negative entries - so that $\pi_{ij}(-\Delta_{ij}) = \pi_{ij}(\Delta_{ij})$. If we only consider positive scalar multiplications - D with all positive entries - then the symmetry condition can also be relaxed.

3.4 Bayesian graphical LASSO review

The GLASSO prior distribution has primarily been explored by Wang [2012] and Khondker et al. [2013]. In this section we will briefly review their work.

A major contribution of the work of Wang [2012] was exploring properties of the GLASSO prior. Through sampling they displayed the marginal densities, after truncation onto the space of positive definite matrices. In particular they showed that for parameter $\lambda = 3$, the marginal densities are more concentrated around 0 than the Laplace density with scale parameter λ^{-1}. More importantly, they showed that for fixed penalty parameter $\lambda = 3$, the marginal densities of the partial correlations become more concentrated around 0 as the dimension p increases. They also claim that the marginal distribution on the partial correlations does not depend on the parameter λ - something we verify in Section 3.5. They also explore a hierarchical representation and show that this representation does indeed correspond to the GLASSO prior.

Another important aspect explored by Wang [2012] is the difference between the posterior mean and MAP estimate. This was done by comparing the two estimates in a specific real data set with $p = 11$, $n = 10$ with the mean being calculated from a Monte Carlo sample. It was shown that the mean and mode can vary significantly in both the diagonal and off-diagonal entries, particularly for small λ. The mode in some cases was even outside of the 95% credible interval centered around the mean. It was also pointed out that this credible interval reduces in width as the parameter λ is increased showing a reduction in estimation uncertainty - something that is not reflected by a point estimate.

Both Wang [2012] and Khondker et al. [2013] go on to propose Monte Carlo sampling schemes for the GLASSO posterior. Wang [2012] proposed a block Gibbs sampler, while Khondker et al. [2013] preferred a random walk Metropolis-Hastings. They also proposed two different methods for graphical model selection from the obtained posterior samples. Khondker et al. [2013] fixed certain elements of Θ to be zero based on credible intervals. Wang [2012] instead approximated the probability $\mathbb{P}(\theta_{ij} = 0)$ under a discrete and continuous mixture prior by comparing the GLASSO posterior mean of θ_{ij} to the posterior mean under some reference prior. If the mean under GLASSO is less than half that under the reference then the edge is not included. As a reference prior Wang [2012] used a conjugate Wishart prior, presumably for computational efficiency and easy computation of the posterior mean. However, in our opinion a more sensible reference prior would be another GLASSO prior, but with parameter value $\lambda = 0$. Indeed, it is admitted by Wang [2012] that this approach is 'ad-hoc and lacks the formal Bayesian interpretation.'

Wang [2012] went on to propose an extension to the GLASSO prior, in a way akin to the adaptive LASSO, which allows different variance parameters on each θ_{ij}.

3.5 GLASSO and PC-GLASSO prior comparison

In this section we compare the GLASSO and PC-GLASSO prior densities. In particular, we investigate the affect of the parameter λ and the dimension p on the marginal densities of the diagonals and partial correlations. Recall that the respective prior densities are given by

$$\pi_G(\Theta) \propto \prod_i \mathrm{Exp}(\theta_{ii}; \lambda/2) \prod_{i<j} \mathrm{Laplace}(\theta_{ij}; 0, \lambda^{-1}) \mathbb{I}(\Theta \in \mathcal{S}),$$

and

$$\pi_{PC}(\theta, \Delta) \propto \prod_i \theta_{ii}^{-2} \prod_{i<j} \mathrm{Laplace}(\Delta_{ij}; 0, \lambda^{-1}) \mathbb{I}(\Delta \in \mathcal{S}_1).$$

We obtained samples from these prior densities using rejection sampling - we sample from the distribution without positive definite truncation, which only requires sampling of independent exponential and Laplace values, and reject the sample if the resulting matrix is not positive definite.

We begin by briefly discussing the affect of increasing the dimension p. As mentioned in Section 3.4, this has already been explored for the GLASSO prior by Wang [2012]. They found that the marginal on both the off-diagonals and the partial correlations becomes more concentrated around 0 as the dimension p increases, and that the marginal on the diagonal entries has larger mean and variance as p increases. These effects of dimension are an obvious side-effect of the truncation onto the space of positive definite matrices, which becomes more restrictive as the dimension p grows.

We found a similar phenomenon with the PC-GLASSO prior; as the dimension p increases, the marginal on the partial correlations becomes more concentrated around 0. However, in this case the marginal on the diagonal entries does not depend on p because the truncation is on the partial correlation matrix Δ.

Now we investigate the affect of the parameter λ on the marginal distributions. To do this we generated samples from both prior distributions for fixed $p = 5$ and $\lambda = 1, 2$ and 4. Figure 3.1 plots the marginal densities of Δ_{12} and θ_{11}. The top left panel verifies the claim of Wang [2012] that the GLASSO prior $\pi_G(\Delta_{ij})$ does not depend on λ, whereas the bottom panel shows that $\pi_G(\theta_{ii})$ is shrunk towards 0 as λ increases. In contrast, the PC-GLASSO prior (top-right panel) on partial correlations $\pi_{PG}(\Delta_{ij})$ concentrates around zero as λ grows. The marginals on the diagonal entries are given by $\pi_{PG}(\theta_{ii}) \propto \theta_{ii}^{-2}$ regardless of λ.

This demonstrates a fundamental difference in how GLASSO and PCGLASSO induce sparsity in the $\theta_{ij} = \Delta_{ij}\sqrt{\theta_{ii}\theta_{jj}}$. PCGLASSO achieves sparsity through regularisation of the partial correlations, while GLASSO does so by shrinking the diagonal θ_{ii}.

3.6 Discussion

In this chapter we have explored two classes of separable priors which assume independence of the parameters before truncation onto the space of positive definite matrices. In regular priors the entries of Θ are separable, while for PC-separable priors the independence is instead placed on the partial correlations. The MAP estimate under such priors is equal to the penalised likelihood estimate under certain regular and PC-separable penalty functions. As such, the results of the previous chapter prove that the MAP estimate under a regular prior does not satisfy scale invariance, while the MAP estimate a symmetric PC-separable prior with $\pi_{ii}(\theta_{ii}) \propto \theta_{ii}^{-c}$ does

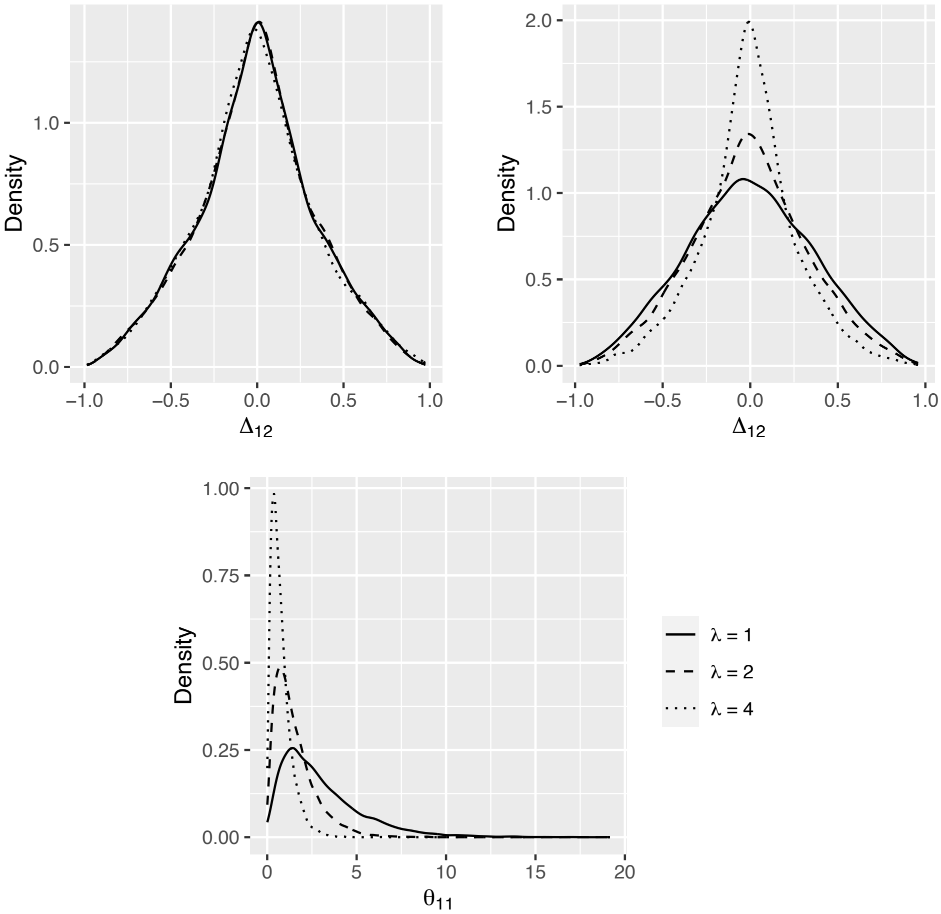


Figure 3.1: Marginal prior densities for the partial correlations under GLASSO prior (top left) and PC-GLASSO prior (top right) and for the diagonal entries under the GLASSO prior (bottom).

satisfy scale invariance. In this chapter we showed that such PC-separable priors in fact satisfy a stronger property of invariance such that any form of posterior inference is invariant to scalar multiplication of the variables.

In this chapter we also proposed a specific PC-separable prior - the PC-GLASSO prior - and compared this to the GLASSO prior. We demonstrated that there are fundamental differences between the two priors and therefore how the corresponding MAP estimates achieve shrinkage and sparsity in the θ_{ij}. In the PC-GLASSO prior, as the parameter λ is increased the marginal on the partial correlations becomes more concentrated around 0, while the marginal on the diagonal entries remains fixed. In the GLASSO prior the opposite is true - the marginal on the partial correlations does not depend on λ, but the marginal on the diagonal entries is shrunk towards 0. In our opinion shrinkage of the partial correlations makes more sense for graphical model selection. The existence of an edge in a graphical model is based on the (partial) dependence between two variables, which can be measured by the partial correlations. The diagonal entries instead relate to the partial variances and are therefore a measure of spread in the data. Within the continuous separable framework, the belief that the underlying graphical model is sparse is therefore best encoded by concentrating the marginal distribution on the partial correlations more tightly around 0.

A related issue is that of prior interpretability. As mentioned earlier in this chapter, separable priors allow great flexibility and simplicity in encoding prior beliefs. For regular priors this involves assigning a distribution to each entry of Θ through π_{ij}. Although these don't exactly correspond to the marginals due to the positive definiteness constraint, prior beliefs may be easily set on the θ_{ij} in this way. However, θ_{ij} is not a good measure of dependence since it is not scale invariant. Thus a large θ_{ij} could reflect either strong dependence between the variables or that they have small partial variance. This is not an ideal situation for setting prior beliefs where one would instead like each parameter to correspond to a single property of the distribution. In this sense, PC-separable priors are advantageous in terms of prior interpretability - one may set prior beliefs on the strength of dependence between variables through $\pi_{ij}(\Delta_{ij})$ and on the spread of each variable through $\pi_{ii}(\theta_{ii})$.

Chapter 4

Spike and slab partial correlation graphical LASSO

In the previous chapter it was mentioned that prior distributions may be used to inspire new penalty functions, and in this chapter we will do exactly that. Spike and slab priors, introduced by Mitchell and Beauchamp [1988] and George and McCulloch [1993], are commonly used in model selection situations where that is some prior probability of a parameter being 0.

Spike and slab distributions for Gaussian graphical models will be more rigorously defined in the next chapter. However, as an introduction, a general spike and slab prior on a parameter ϕ is a mixture of the form

$$\pi(\phi) \propto \eta\pi_1(\phi) + (1-\eta)\pi_0(\phi),$$

where $\eta \in (0,1)$, π_1 is the slab density and π_0 the spike density. The extra parameter η can be considered the probability of some indicator variable γ being equal to 1, $\mathbb{P}(\gamma = 1) = \eta$, $\mathbb{P}(\gamma = 0) = 1 - \eta$. The spike and slab density can then be rewritten as

$$\pi(\phi \mid \gamma) \propto \gamma\pi_1(\phi) + (1-\gamma)\pi_0(\phi).$$

The indicator γ can be related to the parameter ϕ being equal to 0. Most obviously, one might take $\gamma = \mathbb{I}(\phi \neq 0)$ so that η is the prior probability of $\phi \neq 0$. In this case π_0 must be a Dirac mass at 0, while π_1 is the prior density of ϕ conditional on $\phi \neq 0$. Such a formulation was used by Banerjee and Ghosal [2015] in the Gaussian graphical model context. While this is an attractive formulation because parameter selection can be directly related to the posterior distribution of γ, it does come with some difficulties. This is because, the prior distribution (and also the posterior) is not continuous which can make posterior inference and sampling challenging.

A common adaptation to allow easier computation is to let π_0 be a continuous density which is highly concentrated around 0. When this is the case, the indicator variable can be interpreted as $\gamma = 1 - \mathbb{I}(\phi \approx 0)$. Although this formulation no longer allows inference on the event $\{\phi = 0\}$ because the corresponding posterior will always set zero probability to the event, it does come with some positives. First, the prior (and posterior) is now continuous which allows easier computation and posterior inference. Second, it may be argued that inference on this new $\gamma = 1 - \mathbb{I}(\phi \approx 0)$ is beneficial for model selection. If a parameter is known to be suitably close to 0, one may chose to not include this parameter in the model to achieve greater parsimony.

This spike and slab prior framework is completely flexible in the choice of π_1 and π_0 and so they may be chosen to reflect prior beliefs on the parameter. However, when, as we will be doing, a spike and slab prior is used to inspire a penalty function, it is important that $\pi(\phi)$ have a local maximum at 0 and be non-differentiable at 0. This is to ensure that the penalised likelihood is able to achieve exact zero estimates. A popular choice for the spike and slab densities is the Laplace which were used in the spike and slab LASSO of Ročková and George [2018] and the Bayesian Regularization for Graphical Models With Unequal Shrinkage (BAGUS) method of Gan et al. [2018]. For consistency of nomenclature, BAGUS will also be referred to as spike and slab graphical LASSO (SS-GLASSO).

We will begin this chapter by reviewing the work of Gan et al. [2018] on the SS-GLASSO, which utilises a regular prior distribution. We will then propose a new spike and slab method named the spike and slab partial correlation graphical LASSO (SS-PC-GLASSO) which, in fitting with the previous chapters, is based on a PC-separable prior. The corresponding penalty functions obtained from a Laplace spike and slab prior will be compared to other popular penalty functions before a computational method and parameter selection strategy are proposed. The chapter will conclude with some simulation settings and real data applications.

4.1 Spike and slab graphical LASSO

Gan et al. [2018] introduced the following Laplace spike and slab prior on the precision matrix of a Gaussian graphical model

$$\pi(\Theta) \propto \prod_i \pi_{ii}(\theta_{ii}) \prod_{i<j} \pi_{ij}(\theta_{ij}) \mathbb{I}(\Theta \in \mathcal{S}),$$

$$\begin{aligned}\pi_{ii}(\theta_{ii}) &= \tau \exp(-\tau\theta_{ii})\mathbb{I}(\theta_{ii} > 0)\\ &= \mathrm{Exp}(\theta_{ii}; \tau),\end{aligned}$$

$$\begin{aligned}\pi_{ij}(\theta_{ij}) &= \eta \frac{\lambda_1}{2} \exp\left(-\lambda_1 |\theta_{ij}|\right) + (1-\eta)\frac{\lambda_0}{2} \exp\left(-\lambda_0 |\theta_{ij}|\right)\\ &= \eta \, \mathrm{Laplace}(\theta_{ij}; 0, \lambda_1^{-1}) + (1-\eta)\, \mathrm{Laplace}(\theta_{ij}; 0, \lambda_0^{-1})\end{aligned}$$

with $\eta \in (0,1)$ and $\lambda_1 < \lambda_0$ such that the variance of the slab is greater than that of the spike. We refer to this prior as the spike and slab graphical LASSO (SS-GLASSO) prior. It is easy to verify that this is a regular prior distribution with independent priors placed on each entry of Θ before truncating onto the space of positive definite matrices. The prior utilises a spike and slab structure with a Laplace density for both the spike and the slab. The spike density has smaller variance than the slab to ensure that it is more concentrated around 0. The extra parameter η denotes the prior probability of θ_{ij} coming from the slab rather than the spike and can be interpreted as $\eta = 1 - \mathbb{P}(\theta_{ij} \approx 0)$. Note that in this set up the same η is used for all θ_{ij} as well as the same spike and slab densities. This reflects the prior belief that all θ_{ij} are equally likely to be from the slab and that all the dependence relationships between pairs of variables are a priori exchangeable. Because of this symmetry, it is therefore more appropriate to think of η as the prior expected proportion of edges that are from the slab. Similarly to the GLASSO prior, the density π_{ii} on the diagonal entries is Exponential, although in contrast to the GLASSO prior, the parameter τ is different to the parameters of the off-diagonal entries.

Gan et al. [2018] actually make a further truncation onto $||\Theta||_2 \leq B$ for some $B > 0$ where $||\cdot||_2$ is the spectral norm. They went on to show that if $B < \sqrt{2n/\lambda_0}$, then within the region $||\Theta||_2 \leq B$ the posterior distribution will be strictly concave and therefore have a unique maximum. This allows the use of an EM algorithm to find the MAP estimate, which Gan et al. [2018] proposed. This truncation was justified because having a large $||\Theta||_2$ relates to a situation where there are large correlations between the variables - a situation 'in which most methods fail' and in conflict with the assumption of a sparse graphical model. However, this argument ignores the importance of the diagonal entries of Θ - a large spectral norm could occur from large θ_{ii}, not only large correlations between the variables. This gives a further reason, as well as the arguments in Chapter 2, for why standardising the data is vital when implementing the method of Gan et al. [2018].

Gan et al. [2018] only consider the MAP estimate related to the SS-GLASSO prior and so this is equivalent to a penalised likelihood method. This penalty function will be explored further and compared to other non-convex penalties in Section 4.3. However, Gan et al. [2018] utilise the prior interpretation for graphical model selection purposes by considering the additional indicator variables γ_{ij} which indicate whether θ_{ij} comes from the spike or the slab. Posterior inference on the γ_{ij} can then be thought of as graphical model selection. However, in lieu of a full Bayesian analysis, Gan et al. [2018] approximate the posterior edge inclusion probabilities by conditioning on the MAP estimate for Θ and include an edge in the model iff this approximated probability is greater than 0.5. These edge inclusion probabilities seem rather contrived - they are simply a monotone transformation of the estimated $|\theta_{ij}|$ - and do not take into account any posterior uncertainty or dependence between the γ_{ij}. Furthermore, the Laplace priors ensure that exact zero entries are possible, and in fact common, in the MAP estimate meaning that this extra step is not required for model selection - one can simply consider the zero entries in the MAP.

A further important contribution of Gan et al. [2018] were estimation accuracy and selection consistency results for the MAP estimate of the SS-GLASSO prior. Furthermore, these results were shown to extend to certain non-Gaussian distributions satisfying some exponential or polynomial tail condition.

4.2 Spike and slab partial correlation graphical LASSO

As recommended in the previous chapter, we suggest an adaptation of the SS-GLASSO prior which is instead PC-separable and leads to scale invariant posterior inference. This will be of a similar form to the PC-GLASSO prior, however with the Laplace prior on the partial correlations replaced by a Laplace spike and slab. This results in the following spike and slab partial correlation graphical LASSO (SS-PC-GLASSO) prior on the partial correlation matrix Δ and diagonal entries θ:

$$\pi(\theta, \Delta) \propto \prod_i \pi_{ii}(\theta_{ii}) \prod_{i<j} \pi_{ij}(\Delta_{ij}) \mathbb{I}(\Delta \in \mathcal{S}_1),$$

$$\pi_{ii}(\theta_{ii}) = \theta_{ii}^{-2},$$

$$\begin{aligned}\pi_{ij}(\Delta_{ij}) &= \eta \frac{\lambda_1}{2(1-\exp(-\lambda_1))} \exp\left(-\lambda_1 |\Delta_{ij}|\right) + (1-\eta) \frac{\lambda_0}{2(1-\exp(-\lambda_0))} \exp\left(-\lambda_0 |\Delta_{ij}|\right) \\ &= \eta \ \mathrm{TruncLaplace}(\Delta_{ij}; 0, \lambda_1^{-1}) + (1-\eta) \ \mathrm{TruncLaplace}(\Delta_{ij}; 0, \lambda_0^{-1}) \quad (4.1)\end{aligned}$$

with $\eta \in (0,1)$, $\lambda_0 > \lambda_1$ and where TruncLaplace denotes the density of a Laplace distribution truncated onto $(-1,1)$.

The SS-PC-GLASSO prior specifies independent diagonal entries θ_{ii} with density θ_{ii}^{-2}. This ensures, from Proposition 9, that the SS-PC-GLASSO prior leads to scale invariant posterior inference. The partial correlations are treated as independent before truncation onto the space of positive definite matrices and have Laplace spike and slab densities prior to this truncation. It is easy to verify that $\pi_{ij}(\Delta_{ij})$ is decreasing in $|\Delta_{ij}|$ and so the SS-PC-GLASSO prior is symmetric PC-separable.

One important thing to note about the SS-PC-GLASSO prior setup is that the spike and slab densities are *truncated* Laplace densities between -1 and 1. In any spike and slab prior, it is important that the spike and slab densities are proper density functions that integrate to 1. This is because, without this restriction, η will no longer be equal to the prior probability of coming from the slab. For example, in the SS-PC-GLASSO prior, suppose we instead use non-truncated Laplace densities

$$\pi_{ij}(\Delta_{ij}) = \eta \,\mathrm{Laplace}(\Delta_{ij}; 0, \lambda_1^{-1}) + (1-\eta)\,\mathrm{Laplace}(\Delta_{ij}; 0, \lambda_0^{-1}).$$

Here, when considering the restriction $\Delta_{ij} \in (-1,1)$, it is clear that this density will be dominated by the spike. In particular, the prior probability of Δ_{ij} coming from the slab is now equal to

$$\frac{\eta(1-\exp(-\lambda_1))}{\eta(1-\exp(-\lambda_1)) + (1-\eta)(1-\exp(-\lambda_0))}$$

which is smaller than η, and is close to 0 when $\lambda_0 \ll \lambda_1$. To avoid this issue, one must ensure that the spike and slab densities integrate to 1 *after* the $\Delta_{ij} \in (-1,1)$ restriction, as is the case in (4.1).

This discussion has so far ignored the further truncation of both the SS-GLASSO prior and SS-PC-GLASSO prior onto the space of positive definite matrices. The effect of this truncation will be explored further in the Chapter 5.

When combined with the likelihood function, after observing a sample covariance matrix S from n observations, the resulting posterior density is

$$\begin{aligned}
\pi(\theta, \Delta \mid S) &\propto \pi(\theta, \Delta) L(\theta, \Delta \mid S) \\
&\propto \prod_i \theta_{ii}^{-2} \prod_{i<j} \Big(\eta \frac{\lambda_1}{c_1} \exp\left(-\lambda_1 |\Delta_{ij}|\right) + (1-\eta) \frac{\lambda_0}{c_0} \exp\left(-\lambda_0 |\Delta_{ij}|\right) \Big) \\
&\quad \times \det(\Delta)^{\frac{n}{2}} \prod_i \left(\theta_{ii}^{\frac{n}{2}} \right) \exp\left(-\frac{n}{2} \left(\sum_i S_{ii} \theta_{ii} + 2 \sum_{i<j} S_{ij} \sqrt{\theta_{ii}\theta_{jj}} \Delta_{ij} \right) \right) \mathbb{I}(\Delta \in \mathcal{S}_1) \\
&= \det(\Delta)^{\frac{n}{2}} \prod_i \left(\theta_{ii}^{\frac{n-4}{2}} \exp\left(-\frac{n}{2} S_{ii} \theta_{ii} \right) \right) \\
&\quad \times \prod_{i<j} \Big(\exp\left(-n S_{ij} \sqrt{\theta_{ii}\theta_{jj}} \Delta_{ij} \right) \\
&\qquad \times \Big(\eta \frac{\lambda_1}{c_1} \exp\left(-\lambda_1 |\Delta_{ij}|\right) + (1-\eta) \frac{\lambda_0}{c_0} \exp\left(-\lambda_0 |\Delta_{ij}|\right) \Big) \Big) \mathbb{I}(\Delta \in \mathcal{S}_1) \\
&= \det(\Delta)^{\frac{n}{2}} \prod_i \left(\theta_{ii}^{\frac{n-4}{2}} \exp\left(-\frac{n}{2} S_{ii} \theta_{ii} \right) \right) \\
&\quad \times \prod_{i<j} \Big(\Big(\eta \frac{\lambda_1}{c_1} \exp\left(-n S_{ij} \sqrt{\theta_{ii}\theta_{jj}} \Delta_{ij} - \lambda_1 |\Delta_{ij}| \right) \\
&\qquad + (1-\eta) \frac{\lambda_0}{c_0} \exp\left(-n S_{ij} \sqrt{\theta_{ii}\theta_{jj}} \Delta_{ij} - \lambda_0 |\Delta_{ij}| \right) \Big) \Big) \mathbb{I}(\Delta \in \mathcal{S}_1) \qquad (4.2)
\end{aligned}$$

where $c_i = 2(1 - \exp(-\lambda_i))$. This posterior density will be referred to in the computational algorithm of Section 4.4.

4.3 Laplace spike and slab penalty functions

Recall from the previous chapter that the MAP estimate under a prior $\pi(\Theta)$ corresponds to a penalised likelihood estimate with penalty function $Pen(\Theta) = -\log(\pi(\Theta))$. In this section we will investigate the penalty function corresponding to the SS-PC-GLASSO prior and compare this to other popular non-convex penalties.

All prior distributions and penalty functions we discuss in this section will be symmetric PC-separable. As such, a prior density can be written as

$$\pi(\theta, \Delta) \propto \prod_i \pi_{\mathrm{D}}(\theta_{ii}) \prod_{i<j} \pi_{\mathrm{PC}}(\Delta_{ij}) \mathbb{I}(\Delta \in \mathcal{S}_1),$$

and a penalty function as

$$Pen(\theta, \Delta) = \sum_i pen_{\mathrm{D}}(\theta_{ii}) \sum_{i<j} pen_{\mathrm{PC}}(\Delta_{ij}).$$

The correspondence $Pen(\theta, \Delta) = -\log(\pi(\theta, \Delta))$ is therefore obtained by taking

$pen_{\mathrm{D}}(\theta_{ii}) = -\log(\pi_{\mathrm{D}}(\theta_{ii}))$ and $pen_{\mathrm{PC}}(\Delta_{ij}) = -\log(\pi_{\mathrm{PC}}(\Delta_{ij}))$.

As discussed in Sections 2.5 and 3.3, in order for the penalty function or prior distribution to be scale invariant the diagonal penalty must be of the form $pen_{\mathrm{D}}(\theta_{ii}) = c\log(\theta_{ii})$ which corresponds to $\pi_{\mathrm{D}}(\theta_{ii}) = \theta_{ii}^{-c}$ for some $c \geq 0$. This is the case for both PC-GLASSO and SS-PC-GLASSO with the choice of $c = 2$. We recommend that any symmetric PC-separable penalty function or prior distribution have such a diagonal penalty or prior in order to obtain scale invariance.

We now focus on the penalty function on the partial correlations. Two common penalty functions which we will compare to are the L_0 penalty

$$pen_{\mathrm{PC}}(\Delta_{ij}) = \rho\, \mathbb{I}(\Delta_{ij} \neq 0),$$

and the L_1 penalty

$$pen_{\mathrm{PC}}(\Delta_{ij}) = \rho|\Delta_{ij}|,$$

both pictured in Figure 4.1. The L_0 penalty applies the same penalty to all non-zero partial correlations. It is considered by some to be the gold standard for model selection via penalised likelihoods because the penalty is a function of only the model size and not the specific value of the parameters [Dicker et al., 2013]. However, the discontinuity in the penalty functions means that it is not computationally feasible when the model size is even moderately large and can lead to unstable model selection with small changes in the data potentially resulting in large changes to the selected model [Breiman, 1996]. On the other hand, the L_1 penalty, as used in LASSO style methods including the PC-GLASSO, benefits from fast computation due to the convexity of the resulting maximisation problem. However this is often associated with bias in the estimation of large parameter values. This is because the derivative of the L_1 penalty is constant away from 0 and so all parameter values are shrunk towards zero even if the data strongly suggests that one is non-zero. Non-convex penalties can therefore be seen as ways of improving on the problems associated with the L_0 and L_1 penalties - they can be seen as continuous approximations of the L_0 penalty aiding computation or as refinements of the L_1 penalty which reduce the penalty on large parameter values.

The SS-PC-GLASSO prior has

$$\begin{aligned}\pi_{\mathrm{PC}}(\Delta_{ij}) &= \eta\frac{\lambda_1}{2(1-\exp(-\lambda_1))}\exp\left(-\lambda_1|\Delta_{ij}|\right) + (1-\eta)\frac{\lambda_0}{2(1-\exp(-\lambda_0))}\exp\left(-\lambda_0|\Delta_{ij}|\right)\\ &:= \pi_{\mathrm{SS}}(\Delta_{ij}),\end{aligned}$$

which corresponds to the penalty function

$$\begin{aligned} pen_{\mathrm{PC}}(\Delta_{ij}) &= -\log\left(\eta\frac{\lambda_1}{2(1-\exp(-\lambda_1))}\exp\left(-\lambda_1|\Delta_{ij}|\right) + (1-\eta)\frac{\lambda_0}{2(1-\exp(-\lambda_0))}\exp\left(-\lambda_0|\Delta_{ij}|\right)\right) \\ &:= pen_{\mathrm{SS}}(\Delta_{ij}). \end{aligned}$$

One may also wish to consider a version of the SS-PC-GLASSO prior distribution where we set $\lambda_1 = 0$, in which the slab component would correspond to a uniform density between -1 and 1. If this is the case then the first term in $\pi_{\mathrm{SS}}(\Delta_{ij})$ and the corresponding term in $pen_{\mathrm{SS}}(\Delta_{ij})$ should be replaced by their limit as $\lambda_1 \to 0$, which is $\eta/2$.

To better understand the effect of changing parameter values in this penalty function, Figure 4.1 shows plots of pen_{SS} for $\eta = 0.1$ and different values for λ_0, λ_1. Note that the penalty functions have been standardised so that either $pen_{\mathrm{SS}}(0) = 0$ or $pen_{\mathrm{SS}}(1) = 0$ to aid comparison. In the middle left panel we see it plotted for fixed $\lambda_0 = 10$ and $\lambda_1 \in \{0, 2, 5, 10\}$. We see that for $\lambda_0 = \lambda_1$, pen_{SS} is equal to the L_1 penalty. As λ_1 is reduced, the penalty on large partial correlations is reduced, whilst close to zero the penalty remains close to the L_1 penalty.

In the bottom left panel of Figure 4.1 we see pen_{SS} plotted for fixed $\lambda_1 = 0$ and $\lambda_0 \in \{1, 5, 10, 20\}$. For small λ_0 the penalty is close to the L_1 penalty. As λ_0 is increased, the penalty becomes more non-convex and flatter in the extremities. For large λ_0 the penalty begins to resemble a continuous approximation of the L_0 penalty.

More insight into a penalty function can be gained by looking at its derivative. The derivative of the SS-PC-GLASSO penalty is

$$pen'_{\mathrm{SS}}(\Delta_{ij}) = \frac{sign(\Delta_{ij})\left(\eta\frac{\lambda_1^2}{2(1-\exp(-\lambda_1))}\exp\left(-\lambda_1|\Delta_{ij}|\right) + (1-\eta)\frac{\lambda_0^2}{2(1-\exp(-\lambda_0))}\exp\left(-\lambda_0|\Delta_{ij}|\right)\right)}{\eta\frac{\lambda_1}{2(1-\exp(-\lambda_1))}\exp\left(-\lambda_1|\Delta_{ij}|\right) + (1-\eta)\frac{\lambda_0}{2(1-\exp(-\lambda_0))}\exp\left(-\lambda_0|\Delta_{ij}|\right)}$$

Again, if $\lambda_1 = 0$ then the derivative is equal to its limit as $\lambda_1 \to 0$ which is equal to

$$pen'_{\mathrm{SS}}(\Delta_{ij}) = \frac{sign(\Delta_{ij})(1-\eta)\frac{\lambda_0^2}{2(1-\exp(-\lambda_0))}\exp\left(-\lambda_0|\Delta_{ij}|\right)}{\frac{\eta}{2} + (1-\eta)\frac{\lambda_0}{2(1-\exp(-\lambda_0))}\exp\left(-\lambda_0|\Delta_{ij}|\right)}$$

The derivative is often more informative about the dynamics of a penalty function. If the derivative is small or 0 at the MLE then the penalised likelihood estimate will tend to be close to the MLE. If the derivative is large at the MLE then more shrinkage towards 0 can be expected. $pen'_{\mathrm{SS}}(\Delta_{ij})$ is plotted in the middle and bottom right panels of Figure 4.1.

For fixed λ_0 the derivative has a constant value of λ_0, like the L_1 penalty, when $\lambda_1 = \lambda_0$. As λ_1 is decreased, the derivative close to zero is still approximately equal to λ_0, but the derivative decrases with Δ_{ij} even reaching close to 0 for $\lambda_1 = 0$.

For fixed $\lambda_1 = 0$, the derivative close to zero is always approximately equal to λ_0. For small λ_0 the derivative remains approximately constant. As λ_0 is increased, the derivative goes towards 0 for large Δ_{ij} values. For sufficiently large λ_0 the derivative is approximately equal to 0 for all $|\Delta_{ij}|$ above a certain threshold. Note that as the parameter λ_0 increases, the amount of penalisation on large partial correlations increases, but the range of partial correlations for which the penalty function is flat also increases.

A key property of $pen_{\mathrm{SS}}(\Delta_{ij})$ is that it is non-convex. Non-convex penalties have been widely used as a way to reduce the bias in the estimation of large parameter values when using the L_1 penalty. We now compare the SS-PC-GLASSO to two popular non-convex penalties - the Smoothly Clipped Absolute Deviation (SCAD) penalty and the Minimax Concave Penalty (MCP) - on the partial correlations.

The SCAD penalty, proposed by Fan and Li [2001], is symmetric and on $[0, 1)$ is equal to

$$\begin{aligned} pen_{\mathrm{PC}}(\Delta_{ij}) &= \begin{cases} \lambda\Delta_{ij}, & 0 \leq \Delta_{ij} \leq \lambda \\ \frac{2a\lambda\Delta_{ij} - \Delta_{ij}^2 - \lambda^2}{2(a-1)}, & \lambda < \Delta_{ij} \leq a\lambda \\ \frac{1}{2}\lambda^2(a+1), & a\lambda < \Delta_{ij} \end{cases} \\ &:= pen_{\mathrm{SCAD}}(\Delta_{ij}), \end{aligned}$$

which has derivative

$$pen'_{\mathrm{SCAD}}(\Delta_{ij}) = \begin{cases} \lambda, & 0 \leq \Delta_{ij} \leq \lambda \\ \frac{a\lambda - \Delta_{ij}}{(a-1)}, & \lambda < \Delta_{ij} \leq a\lambda \\ 0, & a\lambda < \Delta_{ij} \end{cases}$$

and the related prior density has

$$\begin{aligned} \pi_{\mathrm{PC}}(\Delta_{ij}) &\propto \begin{cases} \exp(-\lambda\Delta_{ij}), & 0 \leq \Delta_{ij} \leq \lambda \\ \exp\left(\frac{-2a\lambda\Delta_{ij} + \Delta_{ij}^2 + \lambda^2}{2(a-1)}\right), & \lambda < \Delta_{ij} \leq a\lambda \\ \exp(-\frac{1}{2}\lambda^2(a+1)), & a\lambda < \Delta_{ij} \end{cases} \\ &:= \pi_{\mathrm{SCAD}}(\Delta_{ij}). \end{aligned}$$

The SCAD penalty contains two parameters - the regularisation parameter λ and an additional parameter a - and is a quadratic spline function with knots at λ and $a\lambda$. Fan and Li [2001] suggested a default value of $a = 3.7$ for the additional parameter. The related prior density π_{SCAD} demonstrates a number of prior beliefs. First, if a partial correlation is small, i.e. $|\Delta_{ij}|\leq \lambda$, then it is likely to be very small. On the other hand, if a partial correlation is large, i.e. $|\Delta_{ij}|> a\lambda$, then the prior is uniform. In between λ and $a\lambda$ the prior is simply a log linear interpolation to ensure continuity between the three components. This gives some interpretation to the parameters a and λ. The parameter λ corresponds to the threshold of partial correlations that are a priori negligible. Meanwhile $a\lambda$ corresponds to the threshold for which any larger partial correlations are significant enough for us to want the data to speak for itself.

The MCP penalty, proposed by Zhang [2010], is also symmetric and on $[0, 1)$ is equal to

$$\begin{aligned} pen_{\text{PC}}(\Delta_{ij}) &= \begin{cases} \lambda\Delta_{ij} - \frac{\Delta_{ij}^2}{2a}, & 0 \leq \Delta_{ij} \leq a\lambda \\ \frac{1}{2}a\lambda^2, & a\lambda < \Delta_{ij} \end{cases} \\ &:= pen_{\text{MCP}}(\Delta_{ij}), \end{aligned}$$

which has derivative

$$pen'_{\text{MCP}}(\Delta_{ij}) = \begin{cases} \lambda - \frac{\Delta_{ij}}{a}, & 0 \leq \Delta_{ij} \leq a\lambda \\ 0, & a\lambda < \Delta_{ij} \end{cases}$$

and the related prior density has

$$\begin{aligned} \pi_{\text{PC}}(\Delta_{ij}) &\propto \begin{cases} \exp\left(-\lambda\Delta_{ij} + \frac{\Delta_{ij}^2}{2a}\right), & 0 \leq \Delta_{ij} \leq a\lambda \\ \exp\left(\frac{1}{2}a\lambda^2\right), & a\lambda < \Delta_{ij} \end{cases} \\ &:= \pi_{\text{MCP}}(\Delta_{ij}). \end{aligned}$$

Like the SCAD penalty, the MCP penalty contains two parameters - the regularisation parameter λ and an additional parameter a - and is a quadratic spline function, but with only a single knot at $a\lambda$. A common default value for the additional parameter is $a = 2$. The MCP prior has a similar interpretation to the SCAD prior with any partial correlations larger than $a\lambda$ being uniform and any smaller than $a\lambda$ likely to be very small.

Plots of the SCAD and MCP penalties and their derivatives can be found in Figure 4.2 for their recommended default value of a and a range of λ values. Both penalty functions seem to act in a similar way with small λ values resulting in more non-convexity in the function whilst for larger λ values the penalties more closely resemble the L_1 penalty.

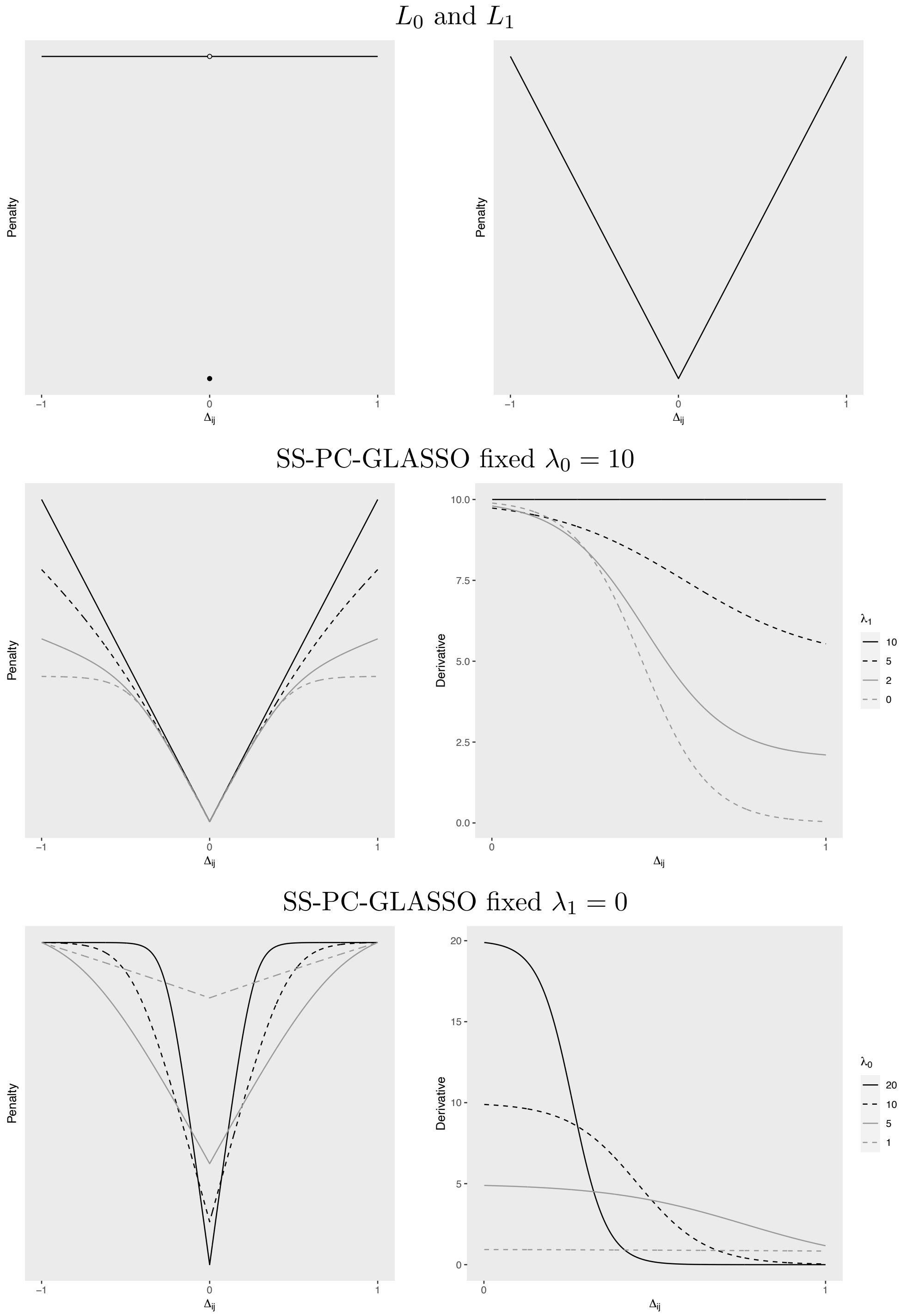
The difference between SCAD and MCP can more easily be seen in their derivatives. Both have piecewise linear derivatives but while the SCAD derivative is constant around 0, the MCP derivative is decreasing around 0. Although not pictured here, changing the additional parameter a has the effect of changing the gradient of the decreasing part of the derivative. This results in a change in the amount of non-convexity in the penalty functions with small a resulting in penalty functions that are closer to the L_0 penalty and large a giving penalty functions closer to the L_1 penalty. This is much the same as the effect of changing λ_1 for fixed λ_0 in the SS-PC-GLASSO. Plots of the SCAD and MCP penalites for different a can be found in Williams [2020].

Comparison of the SS-PC-GLASSO penalty to the SCAD and MCP shows some interesting differences. If the parameter η is treated as fixed in the SS-PC-GLASSO, then all penalty functions have two parameters. Changing λ_1 in the SS-PC-GLASSO has a similar effect to changing the additional parameter a in SCAD and MCP. Meanwhile, λ_0 is more similar to the regularisation parameter λ of SCAD and MCP in that it affects the magnitude of the penalty. However, while SCAD and MCP more closely resemble the L_1 penalty as λ increases, the SS-PC-GLASSO penalty more closely resembles the L_0 penalty as λ_1 increases.

It should be noted that both SCAD and MCP were originally proposed for linear regression and then applied to Gaussian graphical model selection as a regular penalty function. As such they are ordinarily defined over the whole real line rather than just on the interval $(-1, 1)$. Due to this restriction, some of the non-convex properties of these penalties are lost and perhaps alternative default values for the additional parameter a should be considered. For the applications later in this section we consider the regular versions of SCAD and MCP applied on the off-diagonal entries θ_{ij}.

4.4 Parameter selection and computation

Parameter selection for the SS-PC-GLASSO is more complicated than for PC-GLASSO because it contains three parameters - $\lambda_0, \lambda_1, \eta$ - rather than a single parameter. In this section we will propose some strategies for the appropriate selection of these parameters as well as a computational method based on these for obtaining a point estimate for Θ.

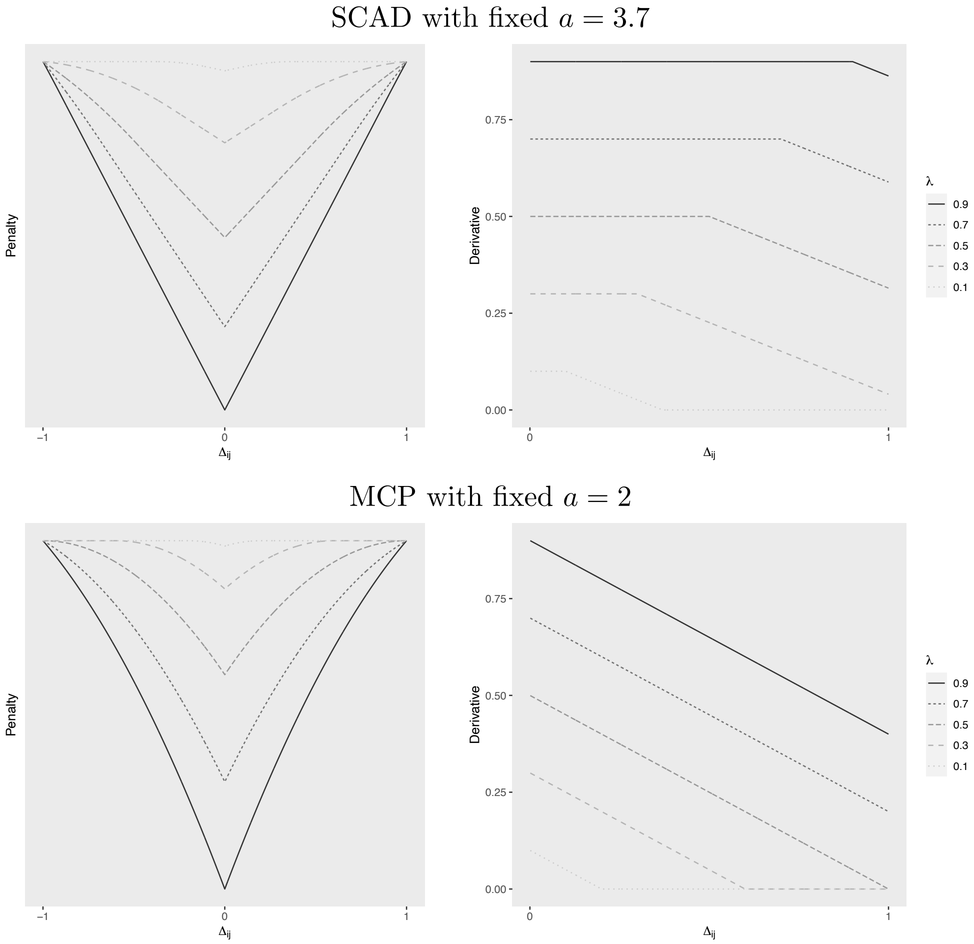


L_0 and L_1

SS-PC-GLASSO fixed $\lambda_0 = 10$

SS-PC-GLASSO fixed $\lambda_1 = 0$

Figure 4.1: Penalty functions (left) and their derivatives (right) (except for top figure which are both penalty functions)



SCAD with fixed $a = 3.7$

MCP with fixed $a = 2$

Figure 4.2: Penalty functions (left) and their derivatives (right)

First we consider the parameter η. Recall that η can be interpreted as the prior probability of a partial correlation coming from the slab rather than the spike. Because this probability is the same for all partial correlations within this prior framework, it is useful to think of η as the prior expected proportion of partial correlations that come from the slab, or the prior expected proportion of edges present in the graphical model. A larger value of η encourages larger graphical models and larger estimated Δ_{ij}, while smaller values of η encourage more sparse models and apply more shrinkage to the Δ_{ij}.

Rather than attempting to tune this parameter, it seems more appropriate to set η based on prior beliefs of the sparsity of the model, or based on the desired level of sparsity in the estimate. Gan et al. [2018] simply fix $\eta = 0.5$ indicating that an edge is equally likely to be present or not. However, we prefer a more conservative approach which encompasses the prior belief that the underlying graphical model is sparse. A common definition of sparsity in a graphical model is that the number of edges is of order p. In fitting with this we propose setting

$$\begin{aligned}\eta &= \frac{p}{\frac{1}{2}p(p-1)} \\ &= \frac{2}{p-1}.\end{aligned}$$

noting that the maximum number of edges in the model is $\frac{1}{2}p(p-1)$.

Next we consider the setting of the spike and slab parameters λ_0 and λ_1. In the PC-GLASSO we proposed setting the single regularisation parameter ρ via the BIC. In this case we were able to consider a large set of values for ρ, which was computationally feasible in Algorithm 1 because the estimate for the previous value of ρ was used as a starting point for the next value of ρ. However, in the SS-PC-GLASSO this is not so straightforward as there are two parameters to consider. Suppose we wished to consider $\lambda_0 \in \{\lambda_0^{(1)}, \ldots, \lambda_0^{(K_0)}\}$ and $\lambda_1 \in \{\lambda_1^{(1)}, \ldots, \lambda_1^{(K_1)}\}$. This will require computation of K_0K_1 estimates. This clearly limits the number of parameters it is feasible to consider, especially when computation is further complicated by the non-convexity of the problem. Regardless, Gan et al. [2018] adopted such an approach, calculating estimates for 16 different combinations of λ_0, λ_1 and selecting the estimate via the BIC.

We propose a different approach in which the SS-PC-GLASSO can be considered a refinement of the PC-GLASSO noting that the two are equivalent when $\lambda_0 = \lambda_1$ by considering the reparameterisation of the PC-GLASSO with $\lambda = n\rho$. We begin by selecting the PC-GLASSO parameter λ via the BIC and fixing this to be the value of the spike parameter λ_0. We then consider a sequence of decreasing slab parameters $\lambda_0 = \lambda_1^{(0)} > \lambda_1^{(1)} > \cdots > \lambda_1^{(k)} = 0$ and again select this parameter via the BIC. This approach, partly inspired by that of Ročková and George [2018], is described in Algorithm 3. The effect of decreasing λ_1 in such a way was seen in Figure 4.1 and can be considered as decreasing the penalty on large partial correlations, in comparison to the L_1 penalty.

In step 3 of Algorithm 3 the spike parameter is chosen by selecting the PC-GLASSO parameter via the BIC. However, the refinement brought about by the slab component of the SS-PC-GLASSO tends to increase the number of edges in the model since it reduces the penalty on large entries. As such, it may be beneficial to select a spike parameter that produces a more sparse model under the PC-GLASSO. This can be achieved, for example, by replacing the BIC in step 3 with the EBIC. In our simulated examples we will consider both the BIC and the EBIC with the additional hyperparameter in the EBIC equal to 0.5 as suggested by Foygel and Drton [2010].

Algorithm 4 describes a coordinate descent algorithm which is used within Algorithm 3. Unlike the previous coordinate descent Algorithm 2, due to the additional complexity of the objective function we update the partial correlations and diagonal entries separately. This algorithm also requires solving two one-dimensional maximisation problems. The first is to maximise the posterior density with respect to a partial correlation Δ_{ij} with all other partial correlations and diagonal entries held fixed. The maximum can easily be found numerically, since Δ_{ij} can only take values on a bounded interval. The next problem is to maximise the posterior with respect to a diagonal entry θ_{ii} with all other diagonal entries and partial correlations held fixed. The solution to this can easily be shown to be equal to

$$\theta_{ii} = \left(\frac{2\left(1 - \frac{4}{n}\right)}{c + \sqrt{c^2 + 4\left(1 - \frac{4}{n}\right) S_{ii}}} \right)^2$$

where

$$c = \sum_{j \neq i} S_{ii} \sqrt{\theta_{ii}} \Delta_{ij}$$

A potential alternative approach would be to fix the slab parameter $\lambda_1 = 0$ so that the slab component is uniform. The spike parameter can then be selected via the BIC in an approach similar to that in Algorithm 1. In the previous section we saw that such an approach would result in a continuous penalty function that closely resembles the L_0 penalty as λ_0 increases. Such an approach would be more directly comparable to the SCAD and MCP penalty functions on the partial correlations. However, the spike and slab interpretation allows the additional parameter to be set in an easily interpretable and principled manner. Both the SCAD and MCP penalties involve an additional parameter with default values for these parameters proposed by Fan and Li [2001] and Zhang [2010] respectively, however without the interpretability that the SS-PC-GLASSO provides. For example, the default parameter value in SCAD was chosen to approximately minimise the Bayes risk under quadratic loss in the linear regression setting, with Bayes risks being computed via numerical integration.

```
Algorithm 3: SS-PC-GLASSO
  Input  : Sample covariance $S$, edge probability parameter $\eta$, sequence
           of parameters $\lambda_0^{(1)} < \cdots < \lambda_0^{(k)}$ and optimisation convergence
           threshold $\epsilon$.
  Output: The MAP estimate $\hat{\Theta}$ for a SS-PC-GLASSO prior, with spike
           parameter $\lambda_0 \in \{\lambda_0^{(1)}, \ldots, \lambda_0^{(k)}\}$ and slab parameter $\lambda_1$ chosen
           via BIC
    1. Standardise the sample covariance $\tilde{S} = \mathrm{diag}(S)^{-1/2} S \mathrm{diag}(S)^{-1/2}$.
    2. Run the PC-GLASSO Algorithm 1 with sample covariance $\tilde{S}$, optimisation
       convergence threshold $\epsilon$ and regularisation parameters $\rho_i = n\lambda_0^{(i)}$,
       $i = 1, \ldots, k$, to obtain a sequence of estimates $\tilde{\Theta}_1, \ldots, \tilde{\Theta}_k$.
    3. Select the estimate from $\tilde{\Theta}_1, \ldots, \tilde{\Theta}_k$ that minimises the BIC, say $\tilde{\Theta}_j$, and set
       $\tilde{\Theta}_{j,0} = \tilde{\Theta}_j$ and $\lambda_0 = \lambda_0^{(j)}$.
    4. Set $0 = \lambda_1^{(k)} < \cdots < \lambda_1^{(1)} < \lambda_0$ such that the $\lambda_1^{(i)}$ are evenly spaced in $[0, \lambda_0)$.
    5. For $i = 1, \ldots, k$, run Algorithm 4 with parameters $\eta, \lambda_0, \lambda_1^{(i)}$, optimisation
       convergence threshold $\epsilon$ and starting point $\tilde{\Theta}_{j,i-1}$ to obtain the estimate $\tilde{\Theta}_{j,i}$.
    6. Select the estimate from $\tilde{\Theta}_{j,0}, \ldots, \tilde{\Theta}_{j,k}$ that minimises the BIC, say $\tilde{\Theta}_{j,i}$, and
       set $\tilde{\Theta} = \tilde{\Theta}_{j,i}$.
    7. Return the estimate $\hat{\Theta} = \mathrm{diag}(S)^{-1/2} \tilde{\Theta} \mathrm{diag}(S)^{-1/2}$.
```

4.5 Applications

We now return to the simulated and real data examples of Section 2.10 to investigate the performance of SS-PC-GLASSO in comparison to PC-GLASSO and other non-convex methods based on regular penalty functions - SCAD, MCP and the

```
Algorithm 4: SS-PC-GLASSO coordinate descent
Input  : Sample covariance $S$ with unit diagonal, parameters $\eta, \lambda_0, \lambda_1$,
         start point $\Theta^{(0)}$ and optimisation convergence threshold $\epsilon$.
Output: A matrix $\Theta$ providing a local maximum of the
         SS-PC-GLASSO posterior density with parameters $\eta, \lambda_0, \lambda_1$.
  1. Let $\Theta^{(1)} = \Theta^{(0)}$ and decompose $\Theta^{(1)}$ to get $\theta^{(1)}$ and $\Delta^{(1)}$.
  2. Cycling randomly without replacement through the set of indices
     $\{(i,j) : i < j; i,j \in \{1,\dots,p\}\}$, do the following:
       (a) Let $\Delta_{ij}$ maximise the posterior density (4.2) subject to $\Delta_{k_1k_2} = \Delta^{(1)}_{k_1k_2}$,
           for all $(k_1,k_2) \neq (i,j)$ and $\theta_{kk} = \theta^{(1)}_{kk}$, for all $k$, and update $\Delta^{(1)}_{ij} = \Delta_{ij}$.
       (b) Let $\theta_{ii}$ maximise the posterior density (4.2) subject to $\Delta_{k_1k_2} = \Delta^{(1)}_{k_1k_2}$,
           for all $(k_1,k_2)$ and $\theta_{kk} = \theta^{(1)}_{kk}$, for all $k \neq i$, and update $\theta^{(1)}_{ii} = \theta_{ii}$.
       (c) Let $\theta_{jj}$ maximise the posterior density (4.2) subject to $\Delta_{k_1k_2} = \Delta^{(1)}_{k_1k_2}$,
           for all $(k_1,k_2)$ and $\theta_{kk} = \theta^{(1)}_{kk}$, for all $k \neq j$, and update $\theta^{(1)}_{jj} = \theta_{jj}$.
  3. Let $q = \max\left\{\frac{2|\{\Delta^{(0)}_{ij} \neq 0 : i<j\}|}{p(p-1)}, \frac{2}{p(p-1)}\right\}$ be the proportion of non-zero
     off-diagonal entries.
  4. If $\pi(\Delta^{(1)}, \theta^{(1)} \mid S)/\pi(\Delta^{(0)}, \theta^{(0)} \mid S) < exp(q\epsilon)$, set $\Delta = \Delta^{(1)}$, $\theta = \theta^{(1)}$ and
     return $\Theta = \theta^{1/2}\Delta\theta^{1/2}$. Otherwise, set $\Delta^{(0)} = \Delta^{(1)}$, $\theta^{(0)} = \theta^{(1)}$ and return to
     Step 2.
```

SS-GLASSO. Two forms of SS-PC-GLASSO are implemented: that in Algorithm 3, and Algorithm 3 with the spike parameter selected in step 3 by the EBIC with parameter 0.5 rather than the BIC. The additional regularisation parameters in SCAD and MCP are set to the default values proposed by Fan and Li [2001] and Zhang [2010] respectively and were implemented using the package **GGMncv** (see Williams [2020]). The BAGUS method which uses a SS-GLASSO prior is implemented using code available online associated to Gan et al. [2018].

4.5.1 Simulations

We consider the same simulated data sets as in Section 2.10 in four different simulation scenarios: the star graph, hub graph, AR2 model and random graph. The methods considered in this section are:

M1. PC-GLASSO

M2. SS-PC-GLASSO with spike parameter selected by the BIC

M3. SS-PC-GLASSO with spike parameter selected by the EBIC with parameter value 0.5

M4. SCAD penalty on data standardised by S

M5. SCAD penalty on data standardised by S^{-1}

M6. MCP penalty on data standardised by S

M7. MCP penalty on data standardised by S^{-1}

M8. BAGUS on data standardised by S

M9. BAGUS on data standardised by S^{-1}

For the SCAD and MCP penalties, estimates are obtained for a long sequence of potential values for the main regularisation parameter. A single estimate is then selected using the BIC.

The results are displayed in Tables 4.1-4.4. We begin by comparing PC-GLASSO to SS-PC-GLASSO. In the star graph, the refinements of SS-PC-GLASSO actually result in a worse estimate than PC-GLASSO in terms of both estimation and model selection. However, this is a setting in which PC-GLASSO performs remarkably well achieving almost perfect model selection even for $n = 30$ and achieving significantly better estimation than other methods tested. In the other three settings, however, SS-PC-GLASSO does offer improvements over PC-GLASSO with either M2 or M3 achieving a better KL loss or MCC in all settings, and often both. These improvements are most notable in the $n = 100$ settings with both M2 and M3 offering large reductions in KL loss when compared to M1.

Between the two SS-PC-GLASSO methods, M3 has better MCC than M2 in all but one of the settings due to its increased specificity. This is to be expected, because use of the EBIC will generally result in a more sparse model and therefore less false positive edges. M3 also has better KL loss than M2 in the star and hub settings, however, M2 generally has better KL loss in the AR2 and random settings. From these results, one may choose to use either M2 or M3 based on whether a high sensitivity or specificity is desired in the particular context. If true edge detection is important then M2 should be preferred. If a more simple model and true non-edge detection is important then M3 should be preffered. However, based on these results we suggest the default method for SS-PC-GLASSO should select the spike parameter via the EBIC as in M3.

Both SCAD and MCP performed poorly in all the $n = 30$ settings, particularly in terms of estimation when data is standardised by S. This, along with SS-PC-GLASSO also offering little advantage over PC-GLASSO in the $n = 30$ settings, gives more evidence for the observation in Section 2.10 that non-convex penalties tend to perform poorly when the sample size n is small. Their performance is much improved in the $n = 100$ settings, with better estimation and model selection than PC-GLASSO in all settings other than the star setting when data is standardised by S. However, this improved performance still generally doesn't match that of SS-PC-GLASSO, which has better estimation in all settings and better model selection in the star and random graphs. This shows that SS-PC-GLASSO improves PC-GLASSO to match or better other non-convex penalties in large sample size settings, whilst maintaining high performance in small sample size settings, unlike SCAD and MCP.

We now compare the SS-PC-GLASSO to the regular SS-GLASSO, or, as it is called in Gan et al. [2018], Bayesian Regularization for Graphical Models With Unequal Shrinkage (BAGUS). In Gan et al. [2018] BAGUS had very good performance in a number of simulation settings in comparison to GLASSO and other competing methods for Gaussian graphical model selection. However, in each simulation setting data was non-standardised and the true underlying Θ had unit diagonal - as discussed in Chapter 2, an idealised scenario for regular penalty functions. Here we have instead standardised the data before applying BAGUS considering both standardisation by S and by S^{-1}. It should also be noted that BAGUS benefits from remarkably fast computation by an EM algorithm with comparable computation to the other non-convex SCAD and MCP methods and not too far from GLASSO, although BAGUS does only consider a small number of potential parameter values. This gives a promising indication that fast computation methods may also be possible for PC-GLASSO and SS-PC-GLASSO.

We also remark that BAGUS is not directly comparable to SS-PC-GLASSO due to the different methods of parameter selection. While SS-PC-GLASSO fixes the spike parameter λ_0 by that selected in PC-GLASSO and considers a range of slab parameters λ_1, BAGUS instead considers a grid of parameter values $\lambda_0 = \tau^{-1} \in (n \log(p))^{1/2} * \{0.05, 0.25, 0.5, 2.5\}$ and $\lambda_1 \in \lambda_0 * \{0.1, 0.2, 1/3, 2/3\}$ with the parameters being selected by the BIC. This grid of parameter values was chosen through theoretical results. It is not immediately obvious how these different strategies might affect the results or if a different strategy may result in improved performance. BAGUS also fixes $\eta = 0.5$ which generally means that it will select less sparse models. This can aid model selection in certain settings by giving increased specificity, but worse selection in others due to decreased sensitivity.

The performance of BAGUS can be highly dependent on the choice of standardisation of the data. For example, in the Star setting the performance of BAGUS is good when data is standardised by S^{-1} but poor under the more common S standardisation. In the AR2 and Random graph settings we see the opposite with model selection being much improved when data is standardised by S.

In Tables 4.1-4.4 we see that the performance of PC-GLASSO and SS-PC-GLASSO is mixed when comparing to BAGUS. In the Star and Hub settings, the partial correlation based methods tended to have the best results for small $n = 30$, whilst BAGUS had better performance for $n = 100$. In the AR2 setting this trend was reversed while in the Random graph setting BAGUS had better estimation and SS-PC-GLASSO had better model selection. In summary, the results of this simulation are not enough to conclude if basing spike and slab priors on partial correlations gives improved performance. Future research in this area may consider identical methods for parameter selection and identical diagonal penalties in order to draw stronger conclusions.

One additional point to consider is sensitivity of SS-PC-GLASSO on the choice of the parameter η. It was pointed out in Ročková and George [2018] that spike and slab methods can be highly sensitive to this choice. Sensitivity to this choice would be reflected in a large number of false positive edges when η is larger than the true proportion of edges in the model, and a large number of false negatives when η is smaller than this proportion. In the simulated examples we used $\eta = \frac{2}{p-1} \approx 0.105$, while the true proportion of edges in the star, hub, AR2 and random models are 0.1, 0.084, 0.195 and 0.147 respectively. Hence η is slightly larger than the true proportion in the star and hub settings, but smaller than the true proportion in the AR2 and random settings. In the star and hub settings SS-PC-GLASSO does not have unusually high numbers of false positives, reflected by the specificity generally being high in comparison to other methods. In the AR2 and random settings, when λ_0 is selected by the BIC, SS-PC-GLASSO actually has large sensitivity in comparison to other methods demonstrating that there is not a large number of false negatives. Hence it seems that, when the λ_0, λ_1 parameters are selected as in SS-PC-GLASSO, the method may not be too sensitive the choice of η. Instead, model selection seems to be driven more by the selection of λ_0.

4.5.2 Gene expression data

Returning to the gene expression application of Section 2.10.2, we investigate how the SS-PC-GLASSO and BAGUS estimates perform in comparison to the PC-

GLASSO, SCAD and MCP in terms of out-of-sample prediction in comparison to model size. The results of this are displayed in Figure 4.3. In this figure the model size vs out-of-sample log-likelihood is plotted for the entire regularisation path for PC-GLASSO, SCAD and MCP, with estimates selected by the BIC displayed by a circle and estimates selected by the EBIC by a square. BAGUS only returns a single estimate rather than a regularisation path and so this estimate is simply displayed by a triangle. Two versions of SS-PC-GLASSO have been considered - that of Algorithm 3 and that of Algorithm 3 with the BIC replaced by the EBIC when selecting the spike parameter λ_0. As such these can be considered refinements of the PC-GLASSO estimates selected by the BIC and EBIC respectively. Model size vs log-likelihood has been plotted for both SS-PC-GLASSO methods for the range of λ_1 values and the estimate selected by the BIC displayed by a circle.

In this example, the SS-PC-GLASSO does not offer improvements over the PC-GLASSO estimate chosen by the BIC. In this case the SS-PC-GLASSO returns a slightly smaller model, but with worse predictive performance - worse than the equivalent model size under the PC-GLASSO. However, when starting at the PC-GLASSO estimate chosen by the EBIC, the SS-PC-GLASSO offers significant improvement. Specifically, the SS-PC-GLASSO has slightly larger model size, but a large increase in the out of sample log-likelihood. This log-likelihood is notably larger than the log-likelihood of the PC-GLASSO estimate of the same size.

The BAGUS estimate on data standardised by S performs very well with a similar model size to the BIC PC-GLASSO estimate but slightly higher log-likelihood. When data is standardised by S^{-1}, however, BAGUS selects a very large model with 6321 edges, significantly larger than any other method. Furthermore, this large model isn't accompanied by a large improvement in log-likelihood meaning this large model is hard to justify. One possible reason for this large model selection is that the parameter values chosen by BAGUS are not suitable for this example.

As discussed in Section 2.10.2, SCAD and MCP perform worse than PC-GLASSO and SS-PC-GLASSO both when data is standardised by S and by S^{-1}.

4.5.3 Stock market data

We now revisit the stock market example of Section 2.10.3. Results are displayed in Figure 4.4. We see that the SS-PC-GLASSO improves upon the PC-GLASSO in both cases where λ_0 is selected by the BIC and EBIC. When selecting the PC-GLASSO parameter via the BIC, the SS-PC-GLASSO improves on this estimate by giving a more sparse estimate with comparable out of sample log-likelihood. When

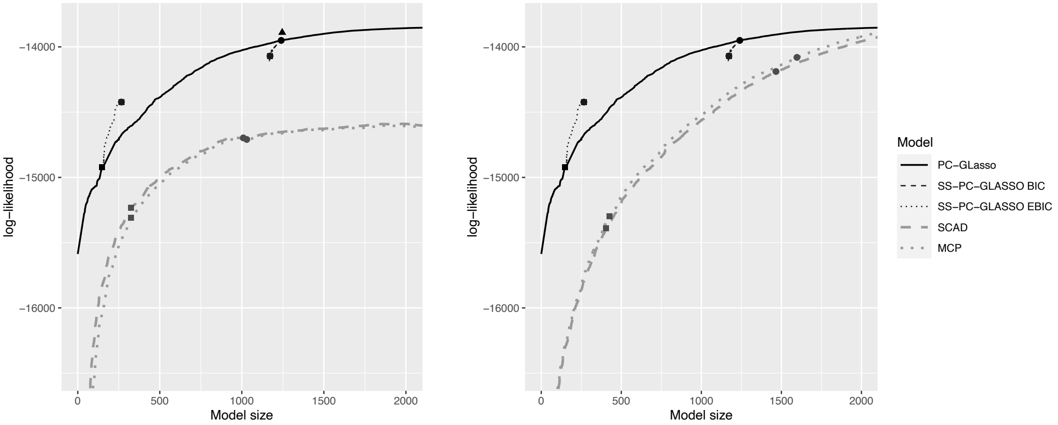


Figure 4.3: Model size vs predictive ability in the gene expression data. Left shows methods on data standardised by S, right shows methods on data standardised by S^{-1}. Estimates selected via BIC and EBIC with $\gamma = 0.5$ are shown by dots and squares respectively. Triangle indicates BAGUS estimate.

selecting the PC-GLASSO parameter via the EBIC, the SS-PC-GLASSO refinement gives an estimate which is of a similar model size but with far greater predictive ability. Furthermore, both SS-PC-GLASSO estimates have better predictive performance compared to the SCAD and MCP estimates of the same model size.

The BAGUS estimate, both when data is standardised by S and by S^{-1}, has a larger model size than other methods with parameters selected by BIC. It also has a similar predictive ability to the PC-GLASSO of the same model size.

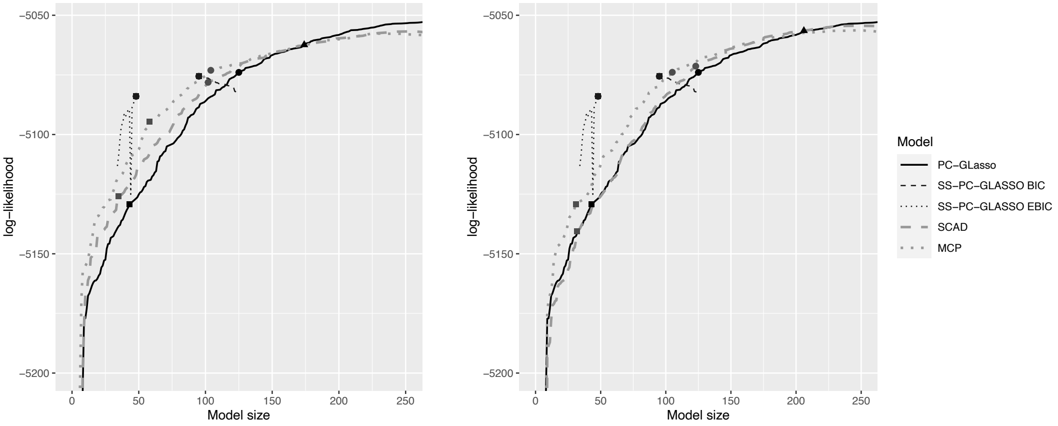


Figure 4.4: Model size vs predictive ability in the stock market data. Left shows methods on data standardised by S, right shows methods on data standardised by S^{-1}. Estimates selected via BIC and EBIC with $\gamma = 0.5$ are shown by dots and squares respectively. Triangle indicates BAGUS estimate.

$n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
M1	**1.42 (0.35)**	**1.69 (0.58)**	**0.978 (0.043)**	0.999 (0.008)	0.995 (0.010)
M2	2.25 (0.65)	2.60 (0.80)	0.874 (0.074)	0.965 (0.040)	0.974 (0.018)
M3	2.07 (0.63)	2.38 (0.77)	0.914 (0.056)	0.971 (0.038)	0.985 (0.011)
M4	8.07 (3.78)	10.87 (4.76)	0.344 (0.136)	0.738 (0.143)	0.764 (0.079)
M5	2.12 (0.99)	2.78 (1.76)	0.527 (0.155)	0.982 (0.037)	0.774 (0.128)
M6	8.58 (4.11)	11.60 (5.17)	0.335 (0.126)	0.737 (0.138)	0.756 (0.079)
M7	2.17 (0.89)	2.71 (1.20)	0.526 (0.148)	0.972 (0.047)	0.783 (0.115)
M8	2.66 (0.59)	4.24 (0.86)	0.259 (0.072)	0.559 (0.098)	0.803 (0.039)
M9	1.59 (0.35)	1.91 (0.58)	0.673 (0.112)	0.994 (0.017)	0.884 (0.063)

$n = 100$	FNorm	KL	MCC	Sensitivity	Specificity
M1	0.70 (0.11)	0.46 (0.12)	0.993 (0.017)	1 (0)	0.999 (0.004)
M2	0.87 (0.18)	0.55 (0.16)	0.959 (0.029)	0.999 (0.007)	0.991 (0.006)
M3	0.85 (0.15)	0.53 (0.13)	0.966 (0.021)	0.999 (0.007)	0.993 (0.004)
M4	1.33 (0.38)	1.01 (0.38)	0.739 (0.135)	0.958 (0.046)	0.926 (0.049)
M5	0.89 (0.13)	0.60 (0.15)	0.734 (0.087)	1 (0)	0.916 (0.038)
M6	1.39 (0.40)	1.09 (0.41)	0.737 (0.128)	0.952 (0.050)	0.928 (0.043)
M7	0.84 (0.13)	0.59 (0.15)	0.837 (0.075)	1 (0)	0.956 (0.025)
M8	1.09 (0.32)	0.80 (0.32)	0.802 (0.139)	0.923 (0.064)	0.959 (0.033)
M9	**0.66 (0.10)**	**0.42 (0.11)**	**1 (0.003)**	1 (0)	1 (0.001)

Table 4.1: Star results

$n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
M1	1.85 (0.29)	2.83 (0.74)	0.696 (0.081)	0.988 (0.043)	0.917 (0.034)
M2	1.96 (0.49)	2.56 (0.82)	0.640 (0.081)	0.968 (0.046)	0.898 (0.035
M3	**1.72 (0.37)**	**2.13 (0.68)**	**0.779 (0.059)**	0.968 (0.054)	0.953 (0.016)
M4	7.80 (4.43)	11.55 (6.33)	0.339 (0.110)	0.830 (0.108)	0.715 (0.115)
M5	2.60 (1.37)	4.04 (2.18)	0.401 (0.098)	0.997 (0.014)	0.675 (0.129)
M6	8.22 (4.68)	12.30 (6.64)	0.329 (0.111)	0.821 (0.112)	0.707 (0.125)
M7	2.54 (1.68)	3.87 (2.69)	0.420 (0.092)	0.994 (0.022)	0.704 (0.115)
M8	1.89 (0.30)	2.44 (0.59)	0.702 (0.067)	0.952 (0.049)	0.930 (0.022)
M9	1.84 (0.23)	2.66 (0.60)	0.533 (0.068)	0.994 (0.023)	0.821 (0.051)

$n = 100$	FNorm	KL	MCC	Sensitivity	Specificity
M1	0.91 (0.15)	0.70 (0.20)	0.858 (0.069)	1 (0)	0.969 (0.019)
M2	0.82 (0.15)	0.53 (0.17)	0.820 (0.059)	1 (0)	0.959 (0.017)
M3	0.80 (0.14)	0.49 (0.15)	0.877 (0.046)	1 (0)	0.975 (0.011)
M4	0.91 (0.21)	0.55 (0.20)	0.918 (0.062)	0.998 (0.012)	0.984 (0.014)
M5	1.05 (0.16)	0.74 (0.15)	0.523 (0.045)	1 (0)	0.814 (0.037)
M6	0.91 (0.22)	0.55 (0.22)	0.920 (0.066)	0.997 (0.014)	0.984 (0.015)
M7	1.05 (0.15)	0.70 (0.16)	0.691 (0.065)	1 (0)	0.912 (0.030)
M8	**0.75 (0.13)**	0.43 (0.11)	0.787 (0.068)	1 (0)	0.948 (0.028)
M9	0.76 (0.13)	**0.41 (0.11)**	**0.934 (0.041)**	1 (0)	0.987 (0.009)

Table 4.2: Hub results

$n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
M1	3.64 (0.31)	5.26 (0.62)	0.283 (0.093)	0.301 (0.194)	0.922 (0.077)
M2	**3.30 (0.54)**	5.63 (1.09)	0.315 (0.070)	0.496 (0.177)	0.830 (0.089)
M3	3.49 (0.26)	5.66 (0.67)	0.289 (0.070)	0.239 (0.075)	0.957 (0.027)
M4	5.98 (4.47)	9.17 (5.56)	0.290 (0.105)	0.444 (0.162)	0.837 (0.114)
M5	4.24 (1.38)	6.46 (2.45)	0.266 (0.085)	0.347 (0.219)	0.869 (0.169)
M6	6.09 (4.61)	9.48 (5.87)	0.270 (0.105)	0.432 (0.159)	0.832 (0.110)
M7	4.15 (2.29)	6.11 (3.33)	0.285 (0.085)	0.451 (0.224)	0.818 (0.175)
M8	3.41 (0.14)	**4.68 (0.48)**	**0.330 (0.066)**	0.494 (0.059)	0.848 (0.029)
M9	3.33 (0.17)	4.86 (0.64)	0.297 (0.066)	0.514 (0.120)	0.808 (0.064)

$n = 100$	FNorm	KL	MCC	Sensitivity	Specificity
M1	2.30 (0.33)	2.00 (0.38)	0.530 (0.052)	0.855 (0.094)	0.774 (0.069)
M2	**1.40 (0.19)**	**1.21 (0.27)**	0.573 (0.062)	0.968 (0.036)	0.736 (0.060)
M3	1.54 (0.24)	1.29 (0.32)	0.728 (0.071)	0.868 (0.070)	0.913 (0.027)
M4	1.60 (0.23)	1.33 (0.29)	0.767 (0.065)	0.908 (0.059)	0.918 (0.039)
M5	1.86 (0.37)	1.45 (0.29)	0.535 (0.049)	0.939 (0.054)	0.720 (0.065)
M6	1.60 (0.23)	1.37 (0.31)	**0.785 (0.065)**	0.895 (0.062)	0.932 (0.035)
M7	1.77 (0.34)	1.37 (0.31)	0.635 (0.059)	0.929 (0.060)	0.817 (0.053)
M8	2.22 (0.19)	1.74 (0.24)	0.632 (0.078)	0.856 (0.063)	0.857 (0.055)
M9	2.16 (0.29)	1.74 (0.54)	0.449 (0.043)	0.907 (0.163)	0.635 (0.117)

Table 4.3: AR2 results

$n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
M1	2.30 (0.25)	3.07 (0.51)	0.336 (0.091)	0.310 (0.153)	0.951 (0.041)
M2	2.59 (0.57)	3.47 (0.88)	0.319 (0.072)	0.420 (0.134)	0.895 (0.052)
M3	2.40 (0.39)	3.27 (0.66)	**0.342 (0.080)**	0.253 (0.079)	0.974 (0.016)
M4	4.87 (4.31)	6.56 (4.81)	0.206 (0.094)	0.318 (0.113)	0.876 (0.078)
M5	2.70 (0.62)	3.89 (0.99)	0.310 (0.080)	0.373 (0.141)	0.908 (0.080)
M6	5.12 (3.83)	6.98 (4.47)	0.194 (0.092)	0.320 (0.112)	0.868 (0.078)
M7	2.46 (0.27)	3.31 (0.57)	0.316 (0.072)	0.427 (0.143)	0.886 (0.075)
M8	**2.08 (0.21)**	**2.65 (0.46)**	0.335 (0.074)	0.509 (0.080)	0.862 (0.029)
M9	2.20 (0.29)	2.97 (0.58)	0.308 (0.084)	0.423 (0.116)	0.886 (0.056)

$n = 100$	FNorm	KL	MCC	Sensitivity	Specificity
M1	1.43 (0.16)	1.23 (0.25)	0.572 (0.059)	0.614 (0.110)	0.941 (0.029)
M2	1.27 (0.15)	0.99 (0.19)	0.570 (0.070)	0.665 (0.088)	0.924 (0.030)
M3	1.28 (0.14)	1.03 (0.20)	**0.623 (0.073)**	0.559 (0.075)	0.975 (0.013)
M4	1.32 (0.15)	1.08 (0.23)	0.598 (0.070)	0.610 (0.105)	0.952 (0.029)
M5	1.37 (0.15)	1.05 (0.19)	0.527 (0.066)	0.701 (0.085)	0.887 (0.036)
M6	1.32 (0.14)	1.09 (0.22)	0.594 (0.070)	0.587 (0.110)	0.957 (0.027)
M7	1.35 (0.16)	1.08 (0.23)	0.580 (0.067)	0.627 (0.100)	0.940 (0.030)
M8	**1.23 (0.12)**	**0.91 (0.16)**	0.534 (0.066)	0.740 (0.069)	0.875 (0.030)
M9	1.28 (0.13)	1.01 (0.21)	0.400 (0.059)	0.795 (0.161)	0.718 (0.095)

Table 4.4: Random graph results

Chapter 5

Spike and slab priors for Gaussian graphical models

In the previous chapter we introduced a new prior distribution, the SS-PC-GLASSO prior, on the precision matrix Θ which is based on a Laplace spike and slab setup on the partial correlations. Because we only considered the MAP estimate, we were able to interpret the prior distribution solely through its related penalty function and the amount of penalisation it placed on non-zero partial correlations. However, when conducting a full posterior analysis using the SS-PC-GLASSO prior, or indeed any separable or PC-separable prior distribution, greater care must be taken to ensure that the beliefs encoded by the prior distribution are as intended.

As a general example of this, consider a $p \times p$ symmetric matrix A with entries a_{ij} and the prior density $\tilde{\pi}$ defined as

$$\tilde{\pi}(A) = \prod_{i \leq j} \pi_{ij}(a_{ij})$$

where π_{ij}, $i, j = 1, \ldots, p$ are some density functions. The prior beliefs encoded by $\tilde{\pi}$ are very simple to interpret - under $\tilde{\pi}$ the entries of A are mutually independent with marginal distributions specified by the π_{ij}.

Now consider the prior density $\tilde{\pi}^+$ on A defined as

$$\begin{aligned} \tilde{\pi}^+(A) &\propto \tilde{\pi}(A)\mathbb{I}(A \in \mathcal{S}) \\ &= \prod_{i \leq j} \pi_{ij}(a_{ij})\mathbb{I}(A \in \mathcal{S}). \end{aligned}$$

The density of $\tilde{\pi}^+$ is exactly the same as $\tilde{\pi}$ except truncated onto the space of positive definite matrices. Under certain additional assumptions on the forms of the π_{ij}, $\tilde{\pi}^+$ can be recognised as a separable prior density from Definition 7. A simple interpretation of $\tilde{\pi}^+$ may be identical to that of $\tilde{\pi}$ - that the entries of A are independent with marginals given by the π_{ij} - with the additional note that it also assumes A to be positive definite. However, it is possible that this is far from the truth. First, under π the entries of A are no longer independent. This can be seen most easily by the restriction $a_{ij} < \sqrt{a_{ii}a_{jj}}$ which is a necessary condition of positive definiteness. Second, the marginal densities of π are no longer given by the π_{ij} and could in fact be quite different. For example, Wang [2012] plotted the marginal distributions for the GLASSO prior which were far from the Exponential and Laplace densities given in the prior setup.

The effect of the truncation can be informally quantified by $\mathbb{P}_{\tilde{\pi}}(A \in \mathcal{S})$ - the probability that A is positive definite under $\tilde{\pi}$. If $\mathbb{P}_{\tilde{\pi}}(A \in \mathcal{S}) \approx 1$ then the truncation will have little affect on $\tilde{\pi}^+$. However, if $\mathbb{P}_{\tilde{\pi}}(A \in \mathcal{S}) \approx 0$, then $\tilde{\pi}^+$ truncates $\tilde{\pi}$ onto a space of infinitesimally small probability and so the resulting prior is vastly different to $\tilde{\pi}$.

Such changes to the marginal distributions can be particularly troublesome when using a spike and slab prior where the spike and slab are used to represent the presence or lack of an edge in the graphical model. The spike and slab framework allows a simple way to encode prior beliefs on the graphical model. However changes to the marginal distributions through the positive definiteness truncation can lead to unintended changes to the prior on the model space. Such an unintended change will be demonstrated in an example later in the section.

Within this chapter we will investigate the effect of the positive definite truncation on spike and slab priors for the precision matrix Θ. We begin in Section 5.1 by defining classes of spike and slab prior distributions on Θ both before and after positive definite truncation. In Section 5.2 we define similar classes of prior distributions but instead based on partial correlations. In Section 5.3 we propose a theorem, based on the theory of Wigner matrices, which determines whether the above probability $\mathbb{P}_{\tilde{\pi}}(A \in \mathcal{S})$ converges to 1 or to 0 as $p \to \infty$. Examples of how this result works in practise is demonstrated on the SS-GLASSO and SS-PC-GLASSO priors introduced in the previous chapter. In Section 5.4 we propose a number of potential choices for spike and slab densities, including a non-local spike and slab, and discuss their respective merits. In Section 5.5 we discuss strategies for performing posterior inference on the model space γ and finish in Section 5.6 with a discussion.

5.1 Regular spike and slab priors

In this section we introduce a class of spike and slab priors based on the off-diagonal entries of Θ. This class of priors will contain the SS-GLASSO prior introduced in the previous chapter and is related to the classes of separable and regular priors introduced in Chapter 3. In order to investigate the effect of the truncation onto the space of positive definite matrices on spike and slab priors, we will begin by defining a class of priors for general symmetric matrices which do not restrict Θ to be positive definite. We then proceed to define a new class of priors by truncating these onto the space of positive definite matrices. To define this class of priors we introduce the random variable $\gamma = \{\gamma_{ij} : 1 \leq i < j \leq p\}$ where each $\gamma_{ij} \in \{0, 1\}$ is an indicator variable.

Definition 10. A prior distribution on (Θ, γ) with density function $\tilde{\pi}(\Theta, \gamma) = \tilde{\pi}(\Theta|\gamma)\tilde{\pi}(\gamma)$ is called a *separable spike and slab (separable-SS) prior* if $\tilde{\pi}(\gamma)$ is any p.m.f. with support on $\{0, 1\}^{|\gamma|}$ and $\tilde{\pi}(\Theta|\gamma)$ can be decomposed as

$$\tilde{\pi}(\Theta|\gamma) = \prod_i \pi_{\mathrm{D}}(\theta_{ii}) \prod_{i<j} \left(\pi_1(\theta_{ij})\mathbb{I}(\gamma_{ij} = 1) + \pi_0(\theta_{ij})\mathbb{I}(\gamma_{ij} = 0)\right) \mathbb{I}(\theta_{ij} = \theta_{ji}), \quad (5.1)$$

where π_{D} is any density on $\mathbb{R}^+$, π_0, π_1 are densities on $\mathbb{R}$ with mean 0 and $\mathrm{Var}_{\pi_0}(\theta_{ij}) < \mathrm{Var}_{\pi_1}(\theta_{ij})$ and $\mathbb{I}$ denotes the indicator function.

If further

$$\tilde{\pi}(\gamma) = \prod_{i<j} \eta^{\gamma_{ij}}(1-\eta)^{1-\gamma_{ij}},$$

for some $\eta \in (0, 1)$, then the prior distribution is called a *regular spike and slab (regular-SS) prior.*

For the remainder of this chapter we will use $\tilde{\pi}$ to denote a general separable-SS prior density and let $\pi_{\mathrm{D}}, \pi_0, \pi_1$ be densities whose role is as given in (5.1).

A separable-SS prior first sets some distribution on the collection of indicator variables γ. Then, conditional on γ, Θ has independent entries, up to symmetry. The diagonal entries each have marginal density π_{D} and θ_{ij} has marginal density π_1 if $\gamma_{ij} = 1$ or π_0 if $\gamma_{ij} = 0$. A regular-SS prior simply adds the condition that the entries of γ are independent and identically distributed with probability η of $\gamma_{ij} = 1$.

The indicator variables γ_{ij} determine if θ_{ij} is marginally distributed according to the spike π_0 or the slab π_1. Setting the spike density π_0 to be a point mass at 0 and the slab density π_1 to be any continuous density leads to the interpretation $\gamma_{ij} = \mathbb{I}(\theta_{ij} \neq 0)$. In this case estimation of γ is exactly equivalent to graphical model selection. However, this point mass tends to cause computational difficulties when calculating the posterior distribution. For this reason, a continuous relaxation of $\theta_{ij} = 0$ is often utilised by allowing π_0 to be a continuous density with small variance, in particular with smaller variance than π_1, see, for example, Ročková and George [2018], Scheipl et al. [2012]. This leads to the interpretation $\gamma_{ij} = 1 - \mathbb{I}(\theta_{ij} \approx 0)$. Although no longer strictly equivalent, inference on γ can be thought of as a proxy for graphical model selection, a practise which is common within spike and slab methods, for example by George and McCulloch [1997]. For this reason we will treat estimation of γ and graphical model selection interchangeably for the remainder of the chapter.

If we let $\eta_{ij} = \mathbb{P}_{\tilde{\pi}}(\gamma_{ij} = 1)$ then the marginal density of θ_{ij} under $\tilde{\pi}$ can be written as

$$\tilde{\pi}(\theta_{ij}) = \eta_{ij}\pi_1(\theta_{ij}) + (1 - \eta_{ij})\pi_0(\theta_{ij}).$$

Although conditionally independent given γ, the entries of Θ are not generally marginally independent under $\tilde{\pi}$. This is because of potential dependence between the γ_{ij} within $\tilde{\pi}(\gamma)$. If in $\tilde{\pi}(\gamma)$ the entries of γ are independent, as is the case for regular-SS priors, then the entries of Θ are also marginally independent under $\tilde{\pi}$. Under a regular-SS prior the marginal density on Θ can be written as

$$\tilde{\pi}(\Theta) = \prod_i \pi_{\mathrm{D}}(\theta_{ii}) \prod_{i<j} (\eta\pi_1(\theta_{ij}) + (1-\eta)\pi_0(\theta_{ij}))\, \mathbb{I}(\theta_{ij} = \theta_{ji}).$$

A separable-SS prior will generally have non-zero probability of Θ not being positive definite and is therefore not appropriate when Θ is a precision matrix. A simple solution to this, as utilised by Wang [2012] and Gan et al. [2018], is to simply truncate such a prior onto the space of positive definite matrices.

Definition 11. A prior distribution with density function $\tilde{\pi}^+(\Theta, \gamma)$ is called a *positive definite separable spike and slab (separable-SS+) prior* if the density can be written as

$$\tilde{\pi}^+(\Theta, \gamma) \propto \tilde{\pi}(\Theta, \gamma)\mathbb{I}(\Theta \in \mathcal{S}),$$

where $\tilde{\pi}$ is a separable-SS prior and $\mathcal{S}$ is the set of symmetric positive definite matrices.

If $\tilde{\pi}$ is also a regular-SS prior then we call $\tilde{\pi}^+$ a *positive definite regular spike and slab (regular-SS+) prior.*

For the remainder of the chapter we use $\tilde{\pi}^+$ to denote the separable-SS+ prior obtained by truncating $\tilde{\pi}$ onto $\mathcal{S}$ and we say that $\tilde{\pi}^+$ and $\tilde{\pi}$ are associated.

If it is further assumed that π_D is non-increasing and that both π_0 and π_1 are non-increasing in $|\theta_{ij}|$ then $\tilde{\pi}^+(\Theta|\gamma)$ is a separable prior density as defined in Definition 7. If $\tilde{\pi}^+$ is also a regular-SS+ prior and π_0, π_1 are continuous densities and symmetric about 0 then the marginal $\tilde{\pi}^+(\Theta)$ is a regular prior distribution. The SS-GLASSO prior introduced by Gan et al. [2018] and in the previous chapter is one example of a regular-SS+ prior which satisfies each of these conditions.

As discussed in the introduction to this chapter, $\tilde{\pi}^+$ can differ greatly from $\tilde{\pi}$ by an amount related to the probability $\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S})$. In particular, the simple interpretation of $\tilde{\pi}$ may no longer be valid for $\tilde{\pi}^+$. Proposition 10 is a trivial observation making this notion precise. We will consider in a later section strategies for ensuring that this probability is close to 1.

Proposition 10. *Let $\tilde{\pi}^+$ be a separable-SS+ prior and $\tilde{\pi}$ be the associated separable-SS prior. Then*

- *The marginal distributions on γ under $\tilde{\pi}^+$ and $\tilde{\pi}$ are related by*

$$\tilde{\pi}^+(\gamma) \propto \tilde{\pi}(\gamma)\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S}|\gamma).$$

- *The conditional distributions of Θ given γ under $\tilde{\pi}^+$ and $\tilde{\pi}$ are related by*

$$\tilde{\pi}^+(\Theta|\gamma) = \frac{\tilde{\pi}(\Theta|\gamma)\mathbb{I}(\Theta \in \mathcal{S})}{\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S}|\gamma)}.$$

Proof. First we show the result for the marginal distribution of γ

$$\begin{aligned}\tilde{\pi}^+(\gamma) &= \int \tilde{\pi}^+(\Theta, \gamma)\, d\Theta \\ &\propto \int \tilde{\pi}(\Theta, \gamma)\mathbb{I}(\Theta \in \mathcal{S})\, d\Theta \\ &= \tilde{\pi}(\gamma) \int \tilde{\pi}(\Theta|\gamma)\mathbb{I}(\Theta \in \mathcal{S})\, d\Theta \\ &= \tilde{\pi}(\gamma)\mathbb{P}(\Theta \in \mathcal{S}|\gamma)\end{aligned}$$

Now we show the result on the conditional distribution of Θ given γ

$$\begin{aligned}\tilde{\pi}^{+}(\Theta|\gamma) &= \frac{\tilde{\pi}^{+}(\Theta,\gamma)}{\tilde{\pi}^{+}(\gamma)} \\ &\propto \frac{\tilde{\pi}(\Theta,\gamma)\mathbb{I}(\Theta \in \mathcal{S})}{\tilde{\pi}(\gamma)\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S}|\gamma)} \\ &\propto \frac{\tilde{\pi}(\Theta|\gamma)\tilde{\pi}(\gamma)\mathbb{I}(\Theta \in \mathcal{S})}{\tilde{\pi}(\gamma)\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S}|\gamma)} \\ &= \frac{\tilde{\pi}(\Theta|\gamma)\mathbb{I}(\Theta \in \mathcal{S})}{\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S}|\gamma)}\end{aligned}$$

□

We further note that under $\tilde{\pi}^{+}$ the entries of Θ are no longer conditionally independent given γ and that in general

$$\tilde{\pi}^{+}(\theta_{ij}|\gamma) \neq \tilde{\pi}^{+}(\theta_{ij}|\gamma_{ij}),$$

meaning that θ_{ij} depends on the whole of γ and not just on γ_{ij}.

An implication of Proposition 10 is that the marginal distribution on the model structure $\tilde{\pi}^{+}(\gamma)$ differs from that in $\tilde{\pi}(\gamma)$ by a factor given by the probability $\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S}|\gamma)$. These probabilities are generally difficult to calculate directly and so it is not straightforward to understand what prior beliefs $\tilde{\pi}^{+}(\gamma)$ imply on the model sparsity. Since

$$\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S}) = \sum_{\gamma} \tilde{\pi}(\gamma)\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S}|\gamma),$$

it is clear that this probability gives a measure of how close $\tilde{\pi}^{+}$ is to $\tilde{\pi}$ overall. Thus, if one wishes to continue using the simple interpretation of $\tilde{\pi}$ to interpret $\tilde{\pi}^{+}$, or embed meaningful prior beliefs, then this probability need be close to 1.

5.2 Partial correlation spike and slab priors

In this section we adapt separable spike and slab priors to instead be separable in the partial correlations. For this we once again use the parameterisation of Θ in terms of the diagonal entries θ and partial correlations Δ. Recall that θ is the diagonal matrix with diagonal entries equal to the diagonal of Θ and Δ is the positive definite symmetric matrix with unit diagonal entries and off diagonal entries $\Delta_{ij} = \frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$. Also recall that Θ is positive definite if and only if Δ is positive definite.

Definition 12. A prior $\pi(\Delta, \theta, \gamma) = \pi(\Delta|\gamma)\pi(\theta)\pi(\gamma)$ is a *partial correlation separable spike and slab (PC-separable-SS) prior* if $\pi(\gamma)$ is any p.m.f. with support on $\{0,1\}^{|\gamma|}$,

$$\pi(\theta) = \prod_i \pi_{\mathrm{D}}(\theta_{ii})$$

where π_{D} is any density on $\mathbb{R}^+$, and $\pi(\Delta|\gamma)$ can be decomposed as

$$\pi(\Delta|\gamma) = \prod_{i<j} \left(\frac{1}{c_1}\pi_1(\Delta_{ij})\mathbb{I}(\gamma_{ij} = 1) + \frac{1}{c_0}\pi_0(\Delta_{ij})\mathbb{I}(\gamma_{ij} = 0) \right) \mathbb{I}(\Delta_{ij}^2 \leq 1)\mathbb{I}(\Delta_{ij} = \Delta_{ji}), \tag{5.2}$$

where π_0, π_1 are densities on $\mathbb{R}$ with mean 0 and $\mathrm{Var}_{\pi_0}(\Delta_{ij}) < \mathrm{Var}_{\pi_1}(\Delta_{ij})$, and

$$c_0 = \int_{-1}^{1} \pi_0(x)\, dx,$$

$$c_1 = \int_{-1}^{1} \pi_1(x)\, dx.$$

If further

$$\pi(\gamma) = \prod_{i<j} \eta^{\gamma_{ij}}(1-\eta)^{1-\gamma_{ij}},$$

for some $\eta \in (0,1)$, then the prior distribution is called a *partial correlation regular spike and slab (PC-regular-SS) prior.*

For the remainder of this chapter we will use π to denote a general PC-separable-SS prior and $\pi_{\mathrm{D}}, \pi_0, \pi_1$ to denote densities satisfying (5.2). Whether $\pi_{\mathrm{D}}, \pi_0, \pi_1$ refer to the separable-SS prior $\tilde{\pi}$ or the PC-separable-SS prior π will be clear from the context.

Under a PC-separable-SS prior the diagonal entries θ_{ii} are independent and identically distributed and, conditional on γ, the partial correlations Δ_{ij} are independent. The marginal distribution of Δ_{ij} conditional on γ only depends on γ through γ_{ij}. If $\gamma_{ij} = 0$ then the marginal density of Δ_{ij} is the spike density

$$\pi(\Delta_{ij}|\gamma_{ij} = 0) = \frac{1}{c_0}\pi_0(\Delta_{ij})\mathbb{I}(\Delta_{ij}^2 \leq 1)$$

and if $\gamma_{ij} = 1$ it is the slab density

$$\pi(\Delta_{ij}|\gamma_{ij} = 1) = \frac{1}{c_1}\pi_1(\Delta_{ij})\mathbb{I}(\Delta_{ij}^2 \leq 1).$$

Notice that the spike and slab densities are both truncated between -1 and 1 with

c_0 and c_1 being the respective normalising constants. This is because $\Delta_{ij} \leq 1$ is a necessary condition for Δ to be positive definite and so truncating the densities in such a way increases the probability of Δ being positive definite under π. One could equivalently restrict π_0 and π_1 to be only defined on $[-1, 1]$. However, the formulation in Definition 12 allows π_0 and π_1 to have the same form as in Definition 10 which aids comparison between separable-SS and PC-separable-SS priors.

Definition 13. A prior distribution with density function $\pi^+(\Delta, \theta, \gamma)$ is called a *positive definite partial correlation separable spike and slab (PC-separable-SS+) prior* if the density can be written as

$$\pi^+(\Delta, \theta, \gamma) \propto \pi(\Delta, \theta, \gamma)\mathbb{I}(\Delta \in \mathcal{S}_1), \tag{5.3}$$

where π is a PC-separable-SS prior and $\mathcal{S}_1$ is the set of symmetric positive definite matrices with unit diagonal.

If π is also a PC-regular-SS prior then we call π^+ a *positive definite partial correlation regular spike and slab (PC-regular-SS+) prior.*

From now on we will use π^+ to denote the PC-separable-SS+ prior obtained by truncating π onto $\Delta \in \mathcal{S}_1$.

An important difference between PC-separable-SS+ priors and separable-SS+ priors is that under the PC-separable-SS+ prior π^+ the diagonal entries θ_{ii} remain independent of each other, Δ and γ. This is because the truncation in (5.3) only involves the partial correlation matrix Δ and not the whole of Θ.

If it is further assumed that π_{D} is non-increasing and that both π_0 and π_1 are non-increasing in $|\Delta_{ij}|$ on $(-1, 1)$ then $\pi^+(\theta, \Delta|\gamma)$ is a PC-separable prior density as defined in Definition 8. If π^+ is also a PC-regular-SS+ prior and π_0, π_1 are symmetric about 0 then the marginal $\pi^+(\theta, \Delta)$ is a regular prior distribution. The SS-PC-GLASSO prior introduced in Section 4.2 is one example of a PC-regular-SS+ prior which satisfies each of these conditions.

When these additional conditions are met, the result of Proposition 9 applies and so any such PC-regular-SS+ prior with $\pi_{\mathrm{D}}(\theta_{ii}) \propto \theta_{ii}^{-c}$ leads to scale invariant posterior inference. This is in contrast to regular-SS+ priors which do not, in general, lead to scale invariant posterior inference. Note also that the proof of Proposition 9 does not rely on the prior density being non-increasing in $|\Delta_{ij}|$. Thus scale invariant posterior inference still holds for PC-regular-SS+ priors even when π_0, π_1 are not non-increasing.

Corollary 4. *Any PC-regular-SS+ prior with $\pi_{\mathrm{D}}(\theta_{ii}) \propto \theta_{ii}^{-c}$ for some $c \geq 0$ and π_0, π_1 symmetric around 0 leads to scale invariant posterior inference.*

An analogous result to Proposition 10 also holds for PC-separable-SS+ priors but with $\Theta \in \mathcal{S}$ being replaced by $\Delta \in \mathcal{S}_1$. We state the result but omit the proof as it is directly analogous to the proof of Proposition 10.

Proposition 11. *Let π^+ be a separable-SS+ prior and π be the associated separable-SS prior. Then*

- *The marginal distributions on γ under π^+ and π are related by*

$$\pi^+(\gamma) \propto \pi(\gamma)\mathbb{P}_\pi(\Delta \in \mathcal{S}_1|\gamma).$$

- *The conditional distributions of Δ given γ under π^+ and π are related by*

$$\pi^+(\Delta|\gamma) = \frac{\pi(\Delta|\gamma)\mathbb{I}(\Delta \in \mathcal{S}_1)}{\mathbb{P}_\pi(\Delta \in \mathcal{S}_1|\gamma)}.$$

Like for separable-SS+ priors, the probability $\mathbb{P}_\pi(\Delta \in \mathcal{S}_1)$ provides a measure of how far π^+ is from π. If this probability is close to 1 then $\pi^+(\gamma) \approx \pi(\gamma)$ and there is only weak dependence between the Δ_{ij} under π^+ with marginal densities similar to those in π. However, if the probability is close to 0 then $\pi^+(\gamma)$ can be far from $\pi(\gamma)$ and the Δ_{ij} can be highly dependent under π^+ with marginals far from those in π.

5.3 Positive definiteness

As discussed in the previous sections, one potential issue with these spike and slab prior frameworks is the truncation onto the space of positive definite matrices. While separable-SS and PC-separable-SS priors are easily interpretable and give a clear framework for setting prior beliefs on the graphical model space through γ, they are not suitable for a precision matrix Θ because they do not impose positive definiteness. Instead, we may choose to truncate these priors onto the space of positive definite matrices in order to obtain a separable-SS+ or PC-separable-SS+ prior. In Propositions 10 and 11 we demonstrated some of the effects the truncation has on the distribution. In particular, the prior on the model space was shown to change by a factor related to the probability $\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S})$ for a separable-SS+ prior and $\mathbb{P}_\pi(\Delta \in \mathcal{S}_1)$ for a PC-separable-SS+ prior. In this section we study these probabilities in the specific case of regular-SS+ and PC-regular-SS+ priors. Our main contribution will be devising an approach that ensures that the probability of positive definiteness under such priors tends to 1 as the dimension p tends to infinity.

This result can inform parameter selection in spike and slab priors in a Bayesian analysis when the number of variables p becomes increasingly large. For smaller p it is usually computationally feasible (and advisable) to use a sampling method to check that the probability of positive definitness is suitably close to 1.

5.3.1 Wigner matrices

The positive definiteness results later in the section rely on the theory of Wigner matrices, which we now briefly review. For more information on the topic of Wigner matrices, see, for example, Anderson et al. [2009].

A Wigner matrix is a type of $p \times p$ symmetric random matrix with independent entries. The diagonal entries and the off-diagonal entries are each identically distributed with finite absolute moments and depend on the matrix dimension p, converging to 0 in probability as $p \to \infty$. More formally, a Wigner matrix is defined as follows.

Definition 14. Let $\{Y_i\}_{1 \leq i}$ and $\{Z_{ij}\}_{1 \leq i < j}$ be two independent families of independent identically distributed, zero mean, real-valued random variables, such that $\mathbb{E}\left[Z_{ij}^2\right] = 1$ and for all integers $k \geq 1$, $r_k := \max\{\mathbb{E}\left[|Z_{ij}|^k\right], \mathbb{E}\left[|Y_i|^k\right]\} < \infty$.

The symmetric $p \times p$ matrix $A_p = (a_{ij})_{1 \leq i,j \leq p}$ with diagonal entries $a_{ii} = \frac{Y_i}{\sqrt{p}}$ and off-diagonal entries $a_{ij} = a_{ji} = \frac{Z_{ij}}{\sqrt{p}}$ is called a *Wigner matrix*.

Much of the theory surrounding Wigner matrices revolves around the distribution of their eigenvalues. A key theorem is that as the matrix dimension $p \to \infty$, the empirical measure of the eigenvalues converges weakly in probability to the so called standard semi-circle distribution which has density function

$$f(x) = \frac{1}{2\pi}\sqrt{4 - x^2}\mathbb{I}(|x| \leq 2).$$

From this it seems clear that the minimal eigenvalue will converge to 2 as $p \to \infty$. However, this property does not in fact hold generally, instead requiring some additional conditions. These are detailed in the following theorem.

Theorem 1. *Let A_p be a $p \times p$ Wigner matrix satisfying $r_k \leq k^{Ck}$ for some constant C and all positive integers k. Then the smallest eigenvalue of A converges in probability to -2 as $p \to \infty$.*

This theorem is adapted from a theorem in Anderson et al. [2009] showing that the largest eigenvalue converges in probability to 2. The result in Theorem 1 is easily proved using either the symmetry of the semi-circle distribution or by simply considering $-A$ which remains a Wigner matrix. A sufficient condition for the additional requirement $r_k \leq k^{Ck}$ is that $|Y_i|$ and $|Z_{ij}|$ posses a finite exponential moment, or that their moment generating functions exist.

5.3.2 Positive definiteness under regular-SS priors

By relating Θ under a regular-SS prior $\tilde{\pi}$ to a Wigner matrix and using the fact that Θ is positive definite if and only if all of it's eigenvalues are positive, we use the result of Theorem 1 to calculate the probability $\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S})$ as $p \to \infty$ for sequences of regular-SS priors.

Proposition 12. *Let $\tilde{\pi}^{(p)}$, $p \geq 1$ be a sequence of regular-SS priors on (Θ_p, γ_p), where Θ_p is a symmetric $p \times p$ matrix, with*

$$\tilde{\pi}^{(p)}(\gamma_p) = \prod_{i<j} \eta_p^{\gamma_{ij}} (1-\eta_p)^{1-\gamma_{ij}},$$

$$\tilde{\pi}^{(p)}(\Theta_p|\gamma_p) = \prod_i \pi_{\mathrm{D}}^{(p)}(\theta_{ii}) \prod_{i<j} \left(\pi_1^{(p)}(\theta_{ij}) \mathbb{I}(\gamma_{ij}=1) + \pi_0^{(p)}(\theta_{ij}) \mathbb{I}(\gamma_{ij}=0) \right).$$

Let $\mu_{\mathrm{D}}^{(p)} = \int x \pi_{\mathrm{D}}^{(p)}(x)\, dx$ be the mean associated with $\pi_{\mathrm{D}}^{(p)}$ and $\sigma_0^{(p)} = \int x^2 \pi_0^{(p)}(x)\, dx$, $\sigma_1^{(p)} = \int x^2 \pi_1^{(p)}(x)\, dx$ be the variances associated with $\pi_0^{(p)}, \pi_1^{(p)}$ respectively, and let $\sigma^{(p)} = \eta_p \sigma_1^{(p)} + (1-\eta_p)\sigma_0^{(p)}$, which is assumed to be finite. Further, assume that the moment generating functions associated to $\pi_{\mathrm{D}}^{(p)}$, $\pi_0^{(p)}$ and $\pi_1^{(p)}$ exist for all p.

(i) If $\lim_{p\to\infty} \frac{\mu_{\mathrm{D}}^{(p)}}{\sqrt{p\, \sigma^{(p)}}} < 2$ then $\lim_{p\to\infty} \mathbb{P}_{\tilde{\pi}^{(p)}}(\Theta_p \in \mathcal{S}) = 0$.

(ii) If $\lim_{p\to\infty} \frac{\mu_{\mathrm{D}}^{(p)}}{\sqrt{p\, \sigma^{(p)}}} > 2$ then $\lim_{p\to\infty} \mathbb{P}_{\tilde{\pi}^{(p)}}(\Theta_p \in \mathcal{S}) = 1$.

Proof. The matrix Θ_p is positive definite if and only if its minimum eigenvalue, which we denote $l_0(\Theta_p)$, is greater than 0.

Under the regular-SS prior $\tilde{\pi}^{(p)}$, the marginal distribution on Θ_p is

$$\tilde{\pi}(\Theta_p) = \prod_i \pi_{\mathrm{D}}(\theta_{ii}) \prod_{i<j} \left(\eta \pi_1(\theta_{ij}) + (1-\eta)\pi_0(\theta_{ij}) \right).$$

Notice that under this density, the entries of Θ_p are independent, up to symmetry, the diagonal entries are identically distributed with density π_{D} and the off-diagonals θ_{ij} are identically distributed with density $\eta\pi_1(\theta_{ij}) + (1-\eta)\pi_0(\theta_{ij})$.

Next consider the matrix $\tilde{\Theta}_p = \frac{1}{\sqrt{p\,\sigma^{(p)}}}\left(\Theta_p - \mu_{\mathrm{D}}^{(p)} I_p\right)$, where I_p denotes the $p \times p$ identity matrix. Notice that $\tilde{\Theta}_p$ is a Wigner matrix under $\tilde{\pi}^{(p)}$ since $\sqrt{p}\tilde{\Theta}_p$ has independent, zero mean entries and unit variance off-diagonal entries. The additional condition giving a bound for the absolute moments in Theorem 1 is also satisfied due to the existence of the moment generating functions for $\pi_{\mathrm{D}}^{(p)}$, $\pi_0^{(p)}$ and $\pi_1^{(p)}$. Also notice that the minimum eigenvalue of $\tilde{\Theta}_p$, $l_0(\tilde{\Theta}_p)$, is related to $l_0(\Theta_p)$ through

$$l_0(\tilde{\Theta}_p) = \frac{1}{\sqrt{p\,\sigma^{(p)}}}\left(l_0(\Theta_p) - \mu_{\mathrm{D}}^{(p)}\right).$$

Hence $l_0(\Theta_p) > 0$ if and only if

$$l_0(\tilde{\Theta}_p) > \frac{-\mu_{\mathrm{D}}^{(p)}}{\sqrt{p\,\sigma^{(p)}}}.$$

Since $\tilde{\Theta}_p$ is a Wigner matrix under $\tilde{\pi}^{(p)}$, and the other conditions of Theorem 1 are met, $l_0(\tilde{\Theta}_p)$ converges in probability to -2. It easily follows that if $\lim_{p\to\infty} \frac{-\mu_{\mathrm{D}}^{(p)}}{\sqrt{p\,\sigma^{(p)}}} > -2$ then $\mathbb{P}_{\tilde{\pi}^{(p)}}\left(l_0(\tilde{\Theta}_p) > \frac{-\mu_{\mathrm{D}}^{(p)}}{\sqrt{p\,\sigma^{(p)}}}\right) \to 0$ as $p \to \infty$ and hence $\mathbb{P}_{\tilde{\pi}^{(p)}}(\Theta_p \in \mathcal{S}) \to 0$.

Similarly, if $\lim_{p\to\infty} \frac{-\mu_{\mathrm{D}}^{(p)}}{\sqrt{p\,\sigma^{(p)}}} < -2$ then $\mathbb{P}_{\tilde{\pi}^{(p)}}(\Theta_p \in \mathcal{S}) \to 1$ as $p \to \infty$.

□

Proposition 12 highlights that the limiting probability of positive definiteness under a sequence of separable-SS priors depends on a simple ratio involving the mean of the diagonal entries, the variance of the off-diagonal entries and the dimension p. Intuitively, the first needs to be sufficiently large relative to the latter. In particular, if $\sqrt{p\,\sigma^{(p)}} \gg \mu_{\mathrm{D}}^{(p)}$ then positive definite matrices receive vanishing probability. So if, for example, we allow $\mu_{\mathrm{D}}^{(p)}$ to remain constant in p, then we require the standard deviation of the off-diagonal entries to be decreasing at a rate quicker than $\frac{1}{\sqrt{p}}$ in order for the probability of positive definiteness to not vanish.

Of course, Proposition 12 only gives a limiting result and is therefore only relevant for large p. For small p it is therefore still important to check the probability of positive definiteness, for example through sampling. We have found in practise that if $\frac{\mu_{\mathrm{D}}^{(p)}}{\sqrt{p\,\sigma^{(p)}}}$ is suitably large then the probability of positive definiteness tends to be close to 1 for finite p.

We now present an example demonstrating the result of Proposition 12 with the SS-GLASSO prior.

Example Recall that the SS-GLASSO prior is a regular-SS+ prior with

$$\begin{aligned}\pi_{\mathrm{D}}(\theta_{ii}) &= \mathrm{Exp}(\theta_{ii}; \tau),\\ \pi_0(\theta_{ij}) &= \mathrm{Laplace}(\theta_{ij}; 0, \lambda_0^{-1}),\\ \pi_1(\theta_{ij}) &= \mathrm{Laplace}(\theta_{ij}; 0, \lambda_1^{-1}).\end{aligned}$$

Hence, under the associated regular-SS prior, the mean of the diagonal entries is τ^{-1} and the variance of the off-diagonal entries is $2\eta\lambda_1^{-2} + 2(1-\eta)\lambda_0^{-2}$.

In the BAGUS method of Gan et al. [2018], a SS-GLASSO prior was used with the following choice of parameter values

$$\begin{aligned}\eta &= 0.5\\ \tau = \lambda_0^{-1} &= c_0\sqrt{\frac{1}{n\log(p)}}\\ \lambda_1^{-1} &= c_1\lambda_0^{-1}\end{aligned}$$

where $c_0 \in \{0.4, 2, 4, 20\}$ and $c_1 \in \{1.5, 3, 5, 10\}$. Note that these prior parameters depend on the data via n. Under the associated regular-SS prior, these choices lead to:

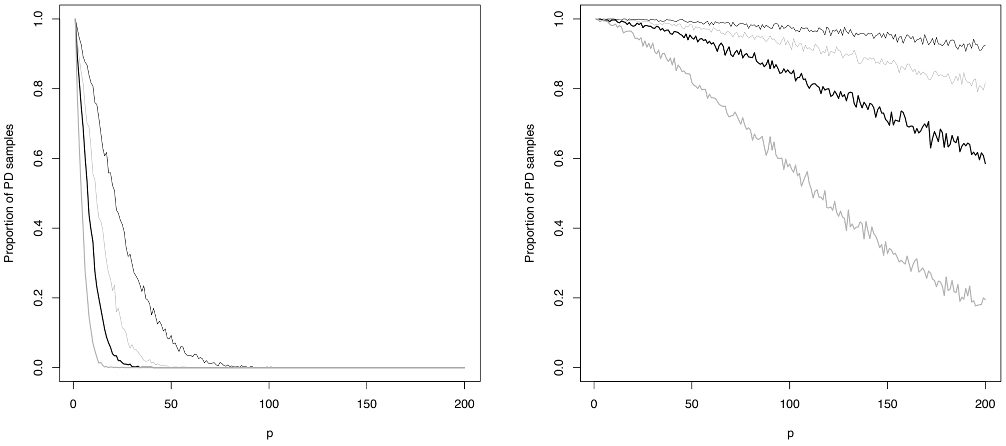
$$\frac{\mu_{\mathrm{D}}^{(p)}}{\sqrt{p\,\sigma^{(p)}}} = \frac{n\log(p)}{c\sqrt{p}}$$

where $c = c_0^2\sqrt{(1+c_1^2)}$. From Prop 1, if $\frac{\sqrt{p}}{\log(p)} \gg n$ (as is the case in high-dimensional settings), then the probability that Θ is positive definite tends to 0 as $p \to \infty$.

We took 1000 samples from the regular-SS prior $\tilde{\pi}$ for $p = 1, \ldots, 100$, fixed $n = 100$ and various values of c_0, c_1 and found the proportion that were positive definite. These proportions can be found in Figure 5.1 for the two lowest values of c_0 and all values of c_1 considered in Gan et al. [2018]. (Note that larger values of c_1 result in a larger variance on the off-diagonals and therefore lower probability of positive definiteness). As predicted by Proposition 12, the proportions tend to decrease with p.

To investigate the affects of Proposition 10 on the related regular-SS+ prior $\tilde{\pi}^+$, we consider the marginal distribution on γ in the case where $p = 10$, $n = 100$, $c_0 = 4$, $c_1 = 10$, under which $\mathbb{P}_{\tilde{\pi}}(\Theta \in \mathcal{S})$ is close to 0. Under $\tilde{\pi}(\gamma)$, the distribution of the number of edges, $|\gamma|$, is binomial with probability $\frac{1}{2}$. However, in Figure 5.2 we see that the distribution of the number of edges under $\tilde{\pi}^+(\gamma)$, obtained via importance sampling, is significantly shifted to the left. This means that the regular-SS+ prior induces more sparsity than specified by the regular-SS prior in a way that

Figure 5.1: The proportion of positive definite samples for $c_0 = 2$ (left) and $c_0 = 0.4$ (right). $c_1 = 10$ (thin black), $c_1 = 5$ (thin grey), $c_1 = 3$ (thick black), $c_1 = 1.5$ (thick grey).



is not easy to control. This therefore restricts the ability to set plausible prior beliefs about the topology of the graph which will drive model selection. Furthermore, this difference is likely to extenuate further when the dimension p increases and the probability $\mathbb{P}(\Theta \in \mathcal{S})$ converges to 0.

5.3.3 Positive definiteness under PC-regular-SS priors

We now present an analogous result for positive definiteness under PC-regular-SS priors.

Proposition 13. *Let $\pi^{(p)}$, $p \geq 1$ be a sequence of PC-regular-SS priors on $(\theta_p, \Delta_p, \gamma_p)$, where θ_p is a $p \times p$ diagonal matrix and Δ_p is a symmetric $p \times p$ matrix, with*

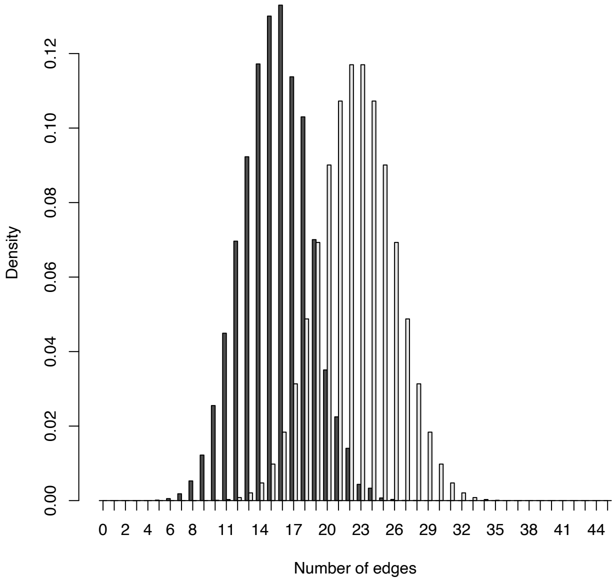
$$\pi^{(p)}(\gamma_p) = \prod_{i<j} \eta_p^{\gamma_{ij}} (1 - \eta_p)^{1-\gamma_{ij}},$$

$$\pi^{(p)}(\theta_p, \Delta_p | \gamma_p) = \prod_i \pi_{\mathrm{D}}^{(p)}(\theta_{ii}) \prod_{i<j} \left(\frac{1}{c_1^{(p)}} \pi_1^{(p)}(\Delta_{ij}) \mathbb{I}(\gamma_{ij} = 1) + \frac{1}{c_0^{(p)}} \pi_0^{(p)}(\Delta_{ij}) \mathbb{I}(\gamma_{ij} = 0) \right) \mathbb{I}(\Delta_{ij}^2 \leq 1).$$

Let $\sigma_0^{(p)} = \frac{1}{c_0^{(p)}} \int_{-1}^{1} x^2 \pi_0^{(p)}(x)\, dx$, $\sigma_1^{(p)} = \frac{1}{c_1^{(p)}} \int_{-1}^{1} x^2 \pi_1^{(p)}(x)\, dx$ be the variances associated with the spike and the slab densities respectively, and let $\sigma^{(p)} = \eta_p \sigma_1^{(p)} + (1 - \eta_p)\sigma_0^{(p)}$, which is assumed to be finite.

(i) If $\lim_{p\to\infty} \frac{1}{\sqrt{p\, \sigma^{(p)}}} < 2$ then $\mathbb{P}_{\pi^{(p)}}(\Delta_p \in \mathcal{S}_1) \to 0$ as $p \to \infty$.

Figure 5.2: The distribution of the number of edges under the regular-SS prior (grey) and regular-SS+ prior (black).



(ii) If $\lim_{p\to\infty} \frac{1}{\sqrt{p\,\sigma^{(p)}}} > 2$ *then* $\mathbb{P}_{\pi^{(p)}}(\Delta_p \in \mathcal{S}_1) \to 1$ *as* $p \to \infty$.

Proof. This proof is very similar to that of Proposition 12. The matrix Δ_p is positive definite if and only if its minimum eigenvalue, which we denote $l_0(\Delta_p)$, is greater than 0.

Under the PC-regular-SS prior $\pi^{(p)}$, the marginal distribution on Δ_p is

$$\tilde{\pi}(\Delta_p) = \prod_{i<j} \left(\frac{\eta}{c_1}\pi_1(\Delta_{ij}) + \frac{(1-\eta)}{c_0}\pi_0(\Delta_{ij}) \right) \mathbb{I}(\Delta_{ij}^2 \leq 1).$$

Notice that under this density, the entries of Δ_p are independent, up to symmetry, the diagonal entries are equal to 1 and the off-diagonals Δ_{ij} are identically distributed with density $\left(\frac{\eta}{c_1}\pi_1(\Delta_{ij}) + \frac{(1-\eta)}{c_0}\pi_0(\Delta_{ij}) \right) \mathbb{I}(\Delta_{ij}^2 \leq 1)$.

Consider the matrix $\tilde{\Delta}_p = \frac{1}{\sqrt{p\,\sigma^{(p)}}}(\Delta_p - I_p)$, where I_p denotes the $p \times p$ identity matrix. Notice that $\tilde{\Delta}_p$ is a Wigner matrix under $\pi^{(p)}$ since $\sqrt{p}\tilde{\Delta}_p$ has independent, zero mean entries and unit variance off-diagonal entries. Also notice that the minimum eigenvalue of $\tilde{\Delta}_p$, $l_0(\tilde{\Delta}_p)$, is related to $l_0(\Delta_p)$ through

$$l_0(\tilde{\Delta}_p) = \frac{1}{\sqrt{p\,\sigma^{(p)}}}\left(l_0(\Delta_p) - 1\right).$$

Hence $l_0(\Delta_p) > 0$ if and only if

$$l_0(\tilde{\Delta}_p) > \frac{-1}{\sqrt{p\,\sigma^{(p)}}}.$$

Since $\tilde{\Delta}_p$ is a Wigner matrix under $\pi^{(p)}$, and the other conditions of Theorem 1 are met (due to the entries of Δ all being in $[-1, 1]$ under $\pi^{(p)}$, $l_0(\tilde{\Delta}_p)$ converges in probability to -2. It easily follows that if $\lim_{p\to\infty} \frac{-1}{\sqrt{p\,\sigma^{(p)}}} > -2$ then $\mathbb{P}_{\pi^{(p)}}\left(l_0(\tilde{\Delta}_p) > \frac{-1}{\sqrt{p\,\sigma^{(p)}}}\right) \to 0$ as $p \to \infty$ and hence $\mathbb{P}_{\pi^{(p)}}(\Delta_p \in \mathcal{S}_1) \to 0$.

Similarly, if $\lim_{p\to\infty} \frac{-1}{\sqrt{p\,\sigma^{(p)}}} < -2$ then $\mathbb{P}_{\pi^{(p)}}(\Delta_p \in \mathcal{S}_1) \to 1$ as $p \to \infty$.

□

Notice two key differences between Propositions 12 and 13. First the limits in Proposition 13 do not depend on the diagonal entry density π_{D}. This is because the condition $\Delta \in \mathcal{S}_1$ does not depend on the diagonal entries θ and under a PC-regular-SS prior, the diagonal entries θ and Δ are independent. This is important as it allows π_{D} to be any density without impacting the probability of positive definiteness.

Second, the restriction of the moment generating functions existing in Proposition 12 is no longer present in Proposition 13. This is because the values of Δ_{ij} are restricted to be in $[-1, 1]$. Hence, under π all absolute moments of Δ_{ij} must be less than or equal to 1. This means that the higher order moment conditions for a Wigner matrix and in Theorem 1 are satisfied for any choice of π_0 and π_1.

We again turn to an example using the SS-PC-GLASSO prior to demonstrate the result of Proposition 13.

Example Recall that the SS-PC-GLASSO prior is a PC-regular-SS+ prior with

$$\begin{aligned}\pi_0(\theta_{ij}) &= \text{Laplace}(\theta_{ij}; 0, \lambda_0^{-1}),\\ \pi_1(\theta_{ij}) &= \text{Laplace}(\theta_{ij}; 0, \lambda_1^{-1}).\end{aligned}$$

We consider the PC-regular-SS prior, π, associated to this. Recall that these densities are truncated between -1 and 1 which has the effect of subtracting a certain amount from the spike and slab variances. Letting $c_i = 1 - \exp(-\lambda_i)$, $i = 0, 1$, the variances associated to the spike, σ_0, and slab, σ_1, are

$$\begin{aligned}\sigma_i &= \frac{1}{c_i}\int_{-1}^{1} x^2 \text{Laplace}(x; 0, \lambda_i^{-1})\, dx \\ &= 2\lambda_i^{-2} - \frac{\lambda_i + 2}{\lambda_i(\exp(\lambda_i) - 1)} \qquad (5.4)\end{aligned}$$

with the variance on the partial correlations being $\sigma = \eta\sigma_1 + (1 - \eta)\sigma_0$. From Proposition 13, this is the key quantity for determining if the probability of positive definiteness converges to 1 or 0.

We now consider a strategy for setting the parameters $\eta, \lambda_0, \lambda_1$, apply Proposition 13 to them and investigate the probability of positive definiteness for finite p using sampling. Recall that it was suggested in Section 4.4 to set $\lambda_1 = 0$. By setting $\lambda_1 = 0$ the truncated Laplace density on the partial correlations becomes a uniform density with variance $\sigma_1 = \frac{1}{3}$. To make calculations easier we consider a limiting case where we allow $\lambda_0 \to \infty$ so that the spike becomes a point mass at 0 with $\sigma_0 = 0$. Hence the variance on the partial correlations becomes $\sigma = \frac{\eta}{3}$ and the result of Proposition 13 depends on the limit of $\sqrt{\frac{3}{p\eta}}$.

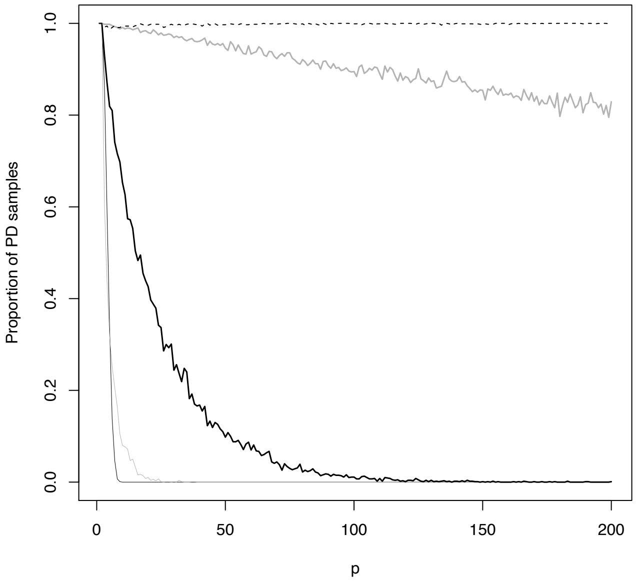
If η is a constant that does not depend on p then clearly $\sqrt{\frac{3}{p\eta}} \to 0$ as $p \to \infty$ and Proposition 13 predicts that the probability of positive definiteness will converge to 0.

In Section 4.4 we proposed allowing η to decrease linearly with p, for example by letting $\eta = \frac{m}{p-1}$. Here, m can be interpreted as the prior expectation of the mean number of edges connected to each vertex. In this case $\sqrt{\frac{3}{p\eta}} \to \sqrt{3/m}$ and so the probability of positive definiteness converges to 0 when $m > \frac{3}{4}$ and to 1 when $m < \frac{3}{4}$.

If η is allowed to converge to 0 at a rate quicker than $\frac{1}{p}$, for example $\eta = \frac{1}{p^2}$, then $\sqrt{\frac{3}{p\eta}} \to \infty$ as $p \to \infty$ and the probability of positive definiteness will converge to 1.

To investigate the probability of positive definiteness in these cases for finite p, we sampled from π 1000 times for each $p = 1, \ldots, 200$ and recorded the proportion of positive definite samples. The results are shown in Figure 5.3. We see that, as expected, the proportion of positive definite samples goes to 0 very quickly in the fixed $\eta = 0.5$ case and for $\eta = \frac{2}{p-1}$. In the case of $\eta = \frac{3}{4(p-1)}$, for which the limiting probability of positive definiteness is not determined by Proposition 13, the proportion of positive definite samples is close to 0 for all $p > 100$. This would suggest that in this borderline case the probability of positive definiteness does still converge to 0. In the case of $\eta = \frac{1}{10(p-1)}$, the proportion of positive definite samples remains high for all p, although away from 1 and generally decreasing with p. However, Proposition 13 predicts that the probability will converge to 1 in this case and so it might be expected that this proportion would eventually increase again for larger p. In the $\eta = \frac{1}{p^2}$ case the observed proportion of positive definite samples is close to 1 for all p.

Figure 5.3: The proportion of positive definite samples for $\eta = 0.5$ (thin black), $\frac{2}{p-1}$ (thin grey), $\frac{3}{4(p-1)}$ (thick black), $\frac{1}{10(p-1)}$ (thick grey), $\frac{1}{p^2}$ (dashed).



This example shows that in order to maintain a high probability of positive definiteness under a PC-regular-SS prior with a uniform slab density, the prior on γ must impose a high level of sparsity. If a prior with less sparsity is desired then one must use a slab density with smaller and shrinking variance as $p \to \infty$.

5.4 Choice of densities

In this section we consider some candidates for the spike and slab densities π_0, π_1 and the diagonal density π_{D}. First, however, we consider the prior on the model space $\pi(\gamma)$.

Under regular-SS and PC-regular-SS priors the prior on the model space has the specific form $\pi(\gamma) = \prod_{i<j} \eta^{\gamma_{ij}}(1-\eta)^{1-\gamma_{ij}}$. That is, the entries of γ are independent, identically distributed Bernoulli random variables with parameter η.

This is a suitable prior when one has prior knowledge of the level of sparsity in the graphical model, but no additional prior information about the structure of the model. In the absence of such additional prior knowledge, we suggest using this $\pi(\gamma)$.

For the the diagonal entries we suggest using a density of the form $\pi_{\mathrm{D}}(\theta_{ii}) \propto \theta_{ii}^{-c}$ for some $c \geq 0$ when dealing with PC-separable-SS priors. This is to ensure scale invariant posterior inference as detailed in Corollary 4.

We now suggest three different possibilities for the spike and slab densities and apply Propositions 12 and 13 to give strategies for setting the parameters to ensure positive definiteness in the limit as $p \to \infty$ under regular-SS and PC-regular-SS priors. First is the Laplace densities of the SS-GLASSO and SS-PC-GLASSO where

$$\pi_i(x) = \text{Laplace}(x; 0, \lambda_i^{-1}).$$

The Laplace is a common choice for spike and slab priors because it is non-differentiable at 0 and therefore the MAP estimate has exact zero entries, as detailed in Chapter 4. Furthermore, in the limits as $\lambda_1 \to 0$ and $\lambda_0 \to \infty$, the slab becomes a uniform density and the spike becomes a point mass at zero, both of which are conceptually quite appealing. In particular, a point mass spike and diffuse heavy-tailed slab are often considered the Bayesian ideal [Castillo and Van Der Vaart, 2012]. The Laplace therefore has the interpretation of being a continuous relaxation of these.

To apply the results of Propositions 12 and 13 we need the variances of the spike and slab densities. In the regular-SS case these variances are $\sigma_i = \int_{-\infty}^{\infty} x^2 \pi_i(x)\, dx = 2\lambda_i^{-2}$. In the PC-regular-SS case these variances are given in (5.4).

Corollary 5. *Let $\tilde{\pi}_{\mathrm{L}}$ and $\tilde{\pi}_{\mathrm{L}}$ be regular-SS and PC-regular-SS priors with Laplace spike and slabs.*

(i) Under $\tilde{\pi}_{\mathrm{L}}$, if $\eta\lambda_1^{-2} + (1-\eta)\lambda_0^{-2} \ll \frac{\mu_{\mathrm{D}}^2}{8p}$, then $\lim_{p\to\infty} \mathbb{P}_{\tilde{\pi}_{\mathrm{L}}}(\Theta_p \in \mathcal{S}) = 1$.

(ii) Under π_{L}, if

$$\eta \left(2\lambda_1^{-2} - \frac{\lambda_1 + 2}{\lambda_1(\exp(\lambda_1) - 1)} \right) + (1-\eta) \left(2\lambda_0^{-2} - \frac{\lambda_0 + 2}{\lambda_0(\exp(\lambda_0) - 1)} \right) \ll \frac{1}{4p},$$

then $\lim_{p\to\infty} \mathbb{P}_{\pi_{\mathrm{L}}}(\Delta_p \in \mathcal{S}_1) = 1$.

Here we use the notation $f(x) \ll g(x)$ to denote $\frac{f(x)}{g(x)} \to 0$ as $x \to \infty$.

From Corollary 5, the variance on the partial correlations must shrink to 0 at a rate faster than $\frac{1}{4p}$ in order for a Laplace PC-regular-SS prior to guarantee positive definiteness as $p \to \infty$.

Since in this chapter we consider the whole posterior distribution, sparsity of the MAP estimate is not required and so we may consider spike and slab densities which are differentiable at 0. The most obvious choice is the Normal density with variance λ_i

$$\pi_i(x) = \mathrm{N}(x; 0, \lambda_i).$$

The advantage of the Normal spike and slab is that its simple form aids computation and sampling from the posterior - see, for example, George and McCulloch [1997]. The Normal spike and slab has been applied to Gaussian graphical models by Wang [2015]. The spike and slab variances in the PC-regular-SS case are given by

$$\begin{aligned}\sigma_i &= \frac{1}{c_i}\int_{-1}^{1} x^2 \pi_i(x)\, dx \\ &= \lambda_i - \frac{2\sqrt{\lambda_i}\phi\left(\frac{1}{\sqrt{\lambda_i}}\right)}{2\Phi\left(\frac{1}{\sqrt{\lambda_i}}\right) - 1}\end{aligned}$$

where $c_i = \int_{-1}^{1} \pi_i(x)\, dx$ and ϕ and Φ are the pdf and cdf of the standard Normal distribution.

Corollary 6. *Let $\tilde{\pi}_{\mathrm{N}}$ and $\tilde{\pi}_{\mathrm{N}}$ be regular-SS and PC-regular-SS priors with Normal spike and slabs.*

(i) Under $\tilde{\pi}_{\mathrm{N}}$, if $\eta\lambda_1 + (1-\eta)\lambda_0 \ll \frac{\mu_{\mathrm{D}}^2}{4p}$, then $\lim_{p\to\infty} \mathbb{P}_{\tilde{\pi}_{\mathrm{N}}}(\Theta_p \in \mathcal{S}) = 1$.

(ii) Under π_{N}, if

$$\eta\left(\lambda_1 - \frac{2\sqrt{\lambda_1}\phi\left(\frac{1}{\sqrt{\lambda_1}}\right)}{2\Phi\left(\frac{1}{\sqrt{\lambda_1}}\right) - 1}\right) + (1-\eta)\left(\lambda_0 - \frac{2\sqrt{\lambda_0}\phi\left(\frac{1}{\sqrt{\lambda_0}}\right)}{2\Phi\left(\frac{1}{\sqrt{\lambda_0}}\right) - 1}\right) \ll \frac{1}{4p},$$

where ϕ and Φ are the pdf and cdf of the standard Normal distribution, then $\lim_{p\to\infty} \mathbb{P}_{\pi_{\mathrm{N}}}(\Delta_p \in \mathcal{S}_1) = 1$.

An interesting alternative to these two choices is to use a non-local density for the slab. In this context, a non-local density is one which is equal to zero at zero, and as such has the attractive property that it assigns zero probability to an edge being present when we condition on the partial correlation being zero. One simple choice for non-local density, introduced by Johnson and Rossell [2010] and shown to have useful properties for Bayesian model selection [Johnson and Rossell, 2012], is the moment (MOM) density:

$$\begin{aligned}\pi_1(x) &= \mathrm{MOM}(x; 0, \lambda_1)\\ &= \frac{x^2}{\lambda_1}\mathrm{N}(\rho; 0, \lambda_1).\end{aligned}$$

The MOM density has variance $3\lambda_1$. When applied to the partial correlations, and therefore truncated onto $(-1, 1)$, the variance is equal to

$$\begin{aligned}\sigma_i &= \frac{1}{c_i}\int_{-1}^{1} x^2\pi_i(x)\,dx\\ &= 3\lambda_1 - \frac{\frac{2}{\sqrt{\lambda_1}}\phi\left(\frac{1}{\sqrt{\lambda_1}}\right)}{2\Phi\left(\frac{1}{\sqrt{\lambda_1}}\right) - \frac{2}{\sqrt{\lambda_1}}\phi\left(\frac{1}{\sqrt{\lambda_1}}\right) - 1}\end{aligned}$$

where $c_i = \int_{-1}^{1}\pi_i(x)\,dx$ and ϕ and Φ still denote the pdf and cdf of the standard Normal distribution.

The MOM slab can be used in conjunction with a Normal spike

$$\pi_0(\rho) = \mathrm{N}(\rho; 0, \lambda_0).$$

We refer to a MOM slab with a Normal spike as simply the MOM spike and slab.

A non-local slab has previously been used in linear regression [Shi et al., 2019], generalised linear models [Bar et al., 2020], and factor regression [Avalos-Pacheco et al., 2020]. For an in depth view of a MOM spike and slab in action, see Avalos Pacheco [2018]. However, to our knowledge non-local priors are yet to be applied to Gaussian graphical models.

Corollary 7. *Let $\tilde{\pi}_{\mathrm{M}}$ and $\tilde{\pi}_{\mathrm{M}}$ be regular-SS and PC-regular-SS priors with MOM spike and slab.*

(i) Under $\tilde{\pi}_{\mathrm{N}}$, if $\eta\lambda_1 + (1-\eta)\lambda_0 \ll \frac{\mu_{\mathrm{D}}^2}{4p}$, then $\lim_{p\to\infty}\mathbb{P}_{\tilde{\pi}_{\mathrm{N}}}(\Theta_p \in \mathcal{S}) = 1$.

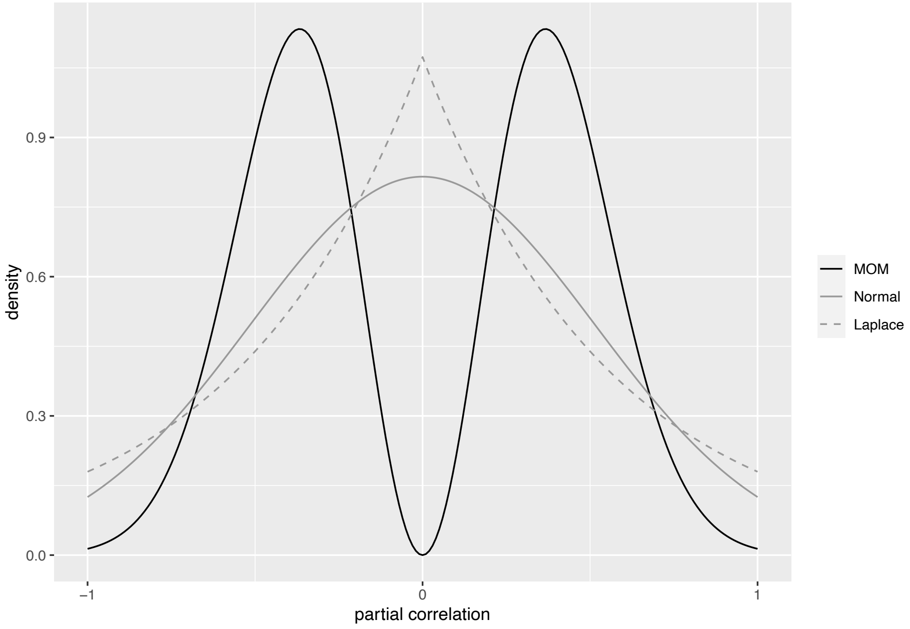
(ii) Under π_{M}, if

$$\eta\left(3\lambda_1 - \frac{\frac{2}{\sqrt{\lambda_1}}\phi\left(\frac{1}{\sqrt{\lambda_1}}\right)}{2\Phi\left(\frac{1}{\sqrt{\lambda_1}}\right) - \frac{2}{\sqrt{\lambda_1}}\phi\left(\frac{1}{\sqrt{\lambda_1}}\right) - 1}\right) + (1-\eta)\left(\lambda_0 - \frac{2\sqrt{\lambda_0}\phi\left(\frac{1}{\sqrt{\lambda_0}}\right)}{2\Phi\left(\frac{1}{\sqrt{\lambda_0}}\right) - 1}\right) \ll \frac{1}{4p},$$

then $\lim_{p\to\infty}\mathbb{P}_{\pi_{\mathrm{L}}}(\Delta_p \in \mathcal{S}_1) = 1$.

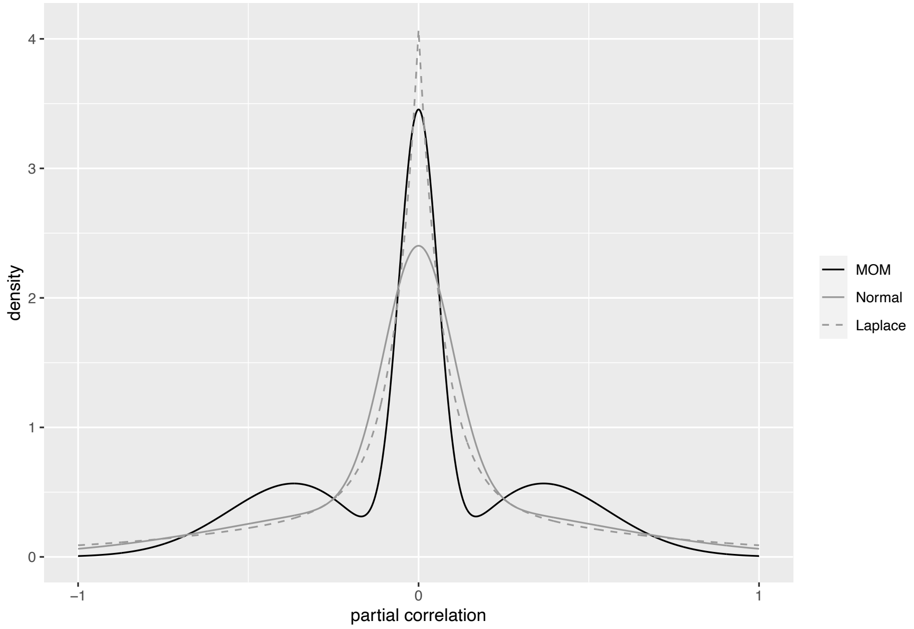
The MOM density is compared with the Normal and Laplaces densities in Figure 5.4. Notice the key feature of the MOM density that it is equal to zero at zero. The density then continuously increases away from zero before reaching two symmetric local maxima. Importantly these local maxima are away from zero and the location of the maxima is controlled by the parameter λ_1 - larger choices of λ_1 lead to these maxima being further from 0. This seems like a good choice for slab density because the slab represents the prior density of the truly non-zero partial correlations. Recall that under a PC-regular-SS prior $\Delta_{ij}|\gamma_{ij} = 1$ follows the slab density π_1. Also recall the interpretation that $\gamma_{ij} = 1 - \mathbb{I}(\Delta_{ij} \approx 0)$. It follows that when $\gamma_{ij} = 1$ then $\Delta_{ij} \neq 0$, something which the MOM density embodies but the Normal and Laplace priors do not. Furthermore, the presence of an edge in a graphical model does not distinguish between positive and negative partial correlations. The MOM density embodies this through symmetry around zero - it encodes that the partial correlation is non-zero but says nothing of its sign.

Figure 5.4: The MOM, Normal and Laplace densities truncated onto $(-1, 1)$ with variance equal to 0.2.



The full spike and slab densities for the Laplace spike and slab, Normal spike and slab and MOM spike and slab are shown in Figure 5.5. The Laplace and Normal spike and slabs look similar to the usual Laplace and Normal densities, however with thicker tails. The MOM spike and slab density is quite different to the other two with the global maximum still at 0 but also with two local maxima - one above zero and one below zero.

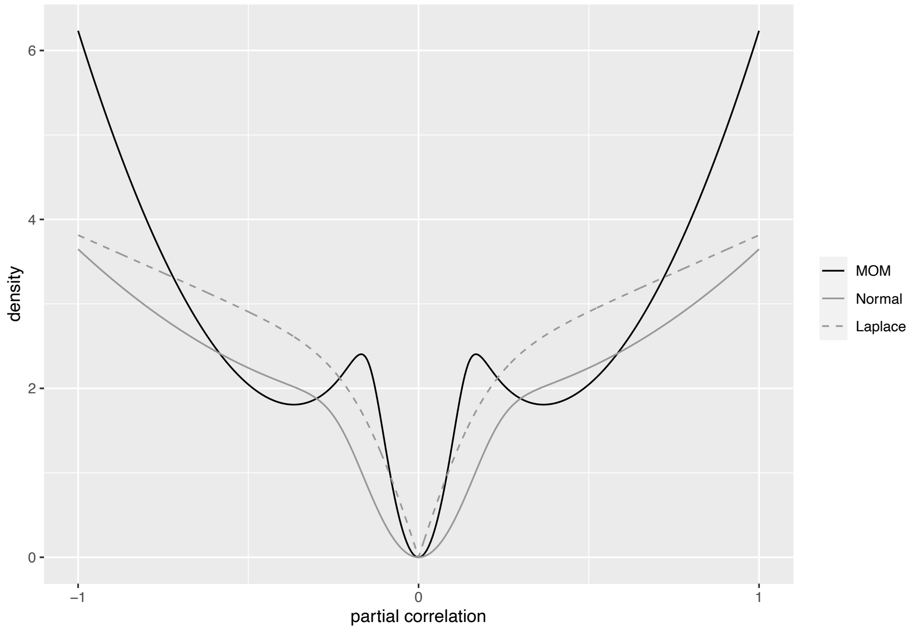
Figure 5.5: The MOM, Normal and Laplace spike and slab densities truncated onto $(-1,1)$ with $\eta = 0.5$ spike variance equal to 0.01 and slab variance equal to 0.2.



In Figure 5.6 we see the associated penalty functions for the three spike and slab priors. The Laplace spike and slab penalty has been discussed in the previous chapter. The Normal spike and slab has a penalty function which has zero derivative at 0 - hence it does not produce zero estimates in the MAP estimate and is not suitable when sparsity in the MAP estimate is desired. The MOM spike and slab has a penalty function very different to those previously discussed because it is not non-decreasing on $\mathbb{R}^+$. This is not standard for penalty functions since large parameter values (in absolute value) are usually penalised more than smaller values. This along with the zero derivative at 0 means that this should not be considered for a penalty function. However, it would still be interesting to investigate how the MAP estimate behaves under the MOM spike and slab in comparison to the Laplace spike and slab.

To summarise, when only considering the MAP estimate or taking a penalised likelihood approach, the Laplace spike and slab, as used in the SS-PC-GLASSO seems the most appropriate of these three choices. However, when considering the whole posterior distribution the Normal spike and slab has provided easier computation when used in other contexts and the MOM spike and slab offers a conceptually appealing choice.

Figure 5.6: The associated penalty functions to the MOM, Normal and Laplace spike and slabs with $\eta = 0.5$ spike variance equal to 0.01 and slab variance equal to 0.2.



5.5 Posterior inference

In this section we consider strategies for posterior inference. The primary strategy we suggest will be based on posterior sampling of Θ. Whilst we don't go as far as proposing a specific algorithm for obtaining such samples, we suggest how such samples may be used to make posterior inference on the model space γ. We also highlight additional considerations that should be made when setting the parameters for the spike and slab densities, particularly in the case of the MOM spike and slab, in light of this strategy.

First we review strategies for posterior inference proposed by others in similar settings. In the spike and slab LASSO in the linear regression setting, Ročková and George [2018] simply found the MAP estimate associated to their Laplace spike and slab (which they treat as a penalty function), and estimate the linear model by the zero coefficient estimates. George and McCulloch [1997], who used the Normal spike and slab in the linear regression setting, proposed a Gibbs sampler which sequentially sampled the regression coefficients, error variance and model indicators γ. The sampled γ were shown to converge in distribution to the posterior on the model space.

In the Gaussian graphical model setting, Banerjee and Ghosal [2015], who proposed a point mass spike and Laplace slab prior, utilised a Laplace approximation in order to obtain approximate posterior model probabilities. Wang [2015] iteratively sampled from the posterior of Θ given γ and γ given Θ under a Normal spike and slab prior.

In the BAGUS method of Gan et al. [2018], in which a regular-SS+ prior with Laplace spike and slab denoted by $\tilde{\pi}^+$ was used, the main strategy for posterior inference was simply finding the MAP estimate $\hat{\Theta}$ under the marginal posterior density $\tilde{\pi}^+(\Theta \mid X)$ via an EM algorithm. Inference on the model space is then conducted by conditioning on the MAP estimate $\tilde{\pi}^+(\gamma \mid \hat{\Theta}, X)$. It was shown that conditional on $\hat{\Theta}$, γ is independent of the data X under $\tilde{\pi}^+$, and further that the entries of γ are independent and γ_{ij} only depends on $\hat{\Theta}$ through $\hat{\theta}_{ij}$. That is,

$$\tilde{\pi}^+(\gamma \mid \hat{\Theta}, X) = \prod_{i<j} \tilde{\pi}^+(\gamma_{ij} \mid \hat{\theta}_{ij}). \tag{5.5}$$

In light of this, posterior inference on γ, conditional on $\Theta = \hat{\Theta}$ being equal to the MAP estimate, is equivalent to finding the probabilities

$$\begin{aligned} \hat{p}_{ij} &= \mathbb{P}_{\tilde{\pi}^+}(\gamma_{ij} = 1 \mid \hat{\theta}_{ij}) \\ &= \frac{\eta \pi_1(\hat{\theta}_{ij})}{\eta \pi_1(\hat{\theta}_{ij}) + (1-\eta)\pi_0(\hat{\theta}_{ij})} \end{aligned} \tag{5.6}$$

Gan et al. [2018] consider $\hat{p}_{ij}$ as an approximation for the posterior probability $p_{ij} = \mathbb{P}_{\tilde{\pi}^+}(\gamma_{ij} = 1 \mid X)$. A single estimate for γ is then obtained by checking if p_{ij} is above or below 0.5, that is $\hat{\gamma}_{ij} = \mathbb{I}(\hat{p}_{ij} > 0.5)$.

While this strategy has its advantages in that it is computationally expedient and returns a single estimated Θ and graphical model, the reliance of the MAP estimate of Θ ignores any posterior uncertainty. Furthermore, there is no guarantee that the estimated $\hat{\gamma}$ is even equal to the MAP estimate of γ maximising $\tilde{\pi}^+(\gamma \mid X)$. However, the properties of the posterior distribution of γ conditional on Θ presented here will be important in our strategy for posterior inference.

Consider a regular-SS+ prior $\tilde{\pi}^+(\Theta, \gamma)$ and suppose we can obtain samples $\Theta^{(1)}, \dots, \Theta^{(K)}$ from the marginal posterior $\tilde{\pi}^+(\Theta \mid X) \propto \tilde{\pi}^+(\Theta) L(\Theta \mid X)$. From here one can easily obtain posterior samples of γ via (5.5) and (5.6). In particular, for $k = 1, \dots, K$, we sample $\gamma^{(k)}$ from $\tilde{\pi}^+(\gamma \mid \Theta^{(k)}, X)$ under which γ has independent entries with $\gamma_{ij} = 1$ with probability

$$\mathbb{P}_{\tilde{\pi}^+}\left(\gamma_{ij} = 1 \mid \theta_{ij}^{(k)}\right) = \frac{\eta\pi_1\left(\theta_{ij}^{(k)}\right)}{\eta\pi_1\left(\theta_{ij}^{(k)}\right) + (1-\eta)\pi_0\left(\theta_{ij}^{(k)}\right)}.$$

The samples $\gamma^{(1)}, \dots, \gamma^{(K)}$ can then be used to estimate posterior properties on γ - for example the MAP model and edge existence probabilities.

For a PC-regular-SS+ prior π^+ a similar strategy can be used with posterior samples $(\theta^{(1)}, \Delta^{(1)}), \dots, (\theta^{(K)}, \Delta^{(K)})$. This time posterior samples for γ can be obtained by sampling the entries of γ independently with $\gamma_{ij} = 1$ with probability

$$\mathbb{P}_{\pi^+}\left(\gamma_{ij} = 1 \mid \Delta_{ij}^{(k)}\right) = \frac{\frac{\eta}{c_1}\pi_1\left(\Delta_{ij}^{(k)}\right)}{\frac{\eta}{c_1}\pi_1\left(\Delta_{ij}^{(k)}\right) + \frac{(1-\eta)}{c_0}\pi_0\left(\Delta_{ij}^{(k)}\right)}.$$

Although we don't propose a method for obtaining such samples of Θ or (θ, Δ) there is reason to believe this to be possible. In the Normal spike and slab, Gibbs sampling has been successfully utilised in the linear regression setting [George and McCulloch, 1993, 1997] and for Gaussian graphical models [Wang, 2015]. The Laplace spike and slab is not too much of a departure from the Bayesian GLASSO prior, for which block Gibbs sampler and random walk Metropolis-Hastings algorithms have been proposed by Wang [2012] and Khondker et al. [2013] respectively. For the MOM spike and slab a Gibbs sampler has been utilised by Shi et al. [2019] for linear regression.

One consideration that this method highlights is that the probabilities $\mathbb{P}_{\tilde{\pi}^+}\left(\gamma_{ij} = 1 \mid \theta_{ij}^{(k)}\right)$, in the case of regular-SS+ priors, and $\mathbb{P}_{\pi^+}\left(\gamma_{ij} = 1 \mid \Delta_{ij}^{(k)}\right)$, in the case of PC-regular-SS+ priors, should be increasing in $|\theta_{ij}|$ and $|\Delta_{ij}|$ respectively. That is, conditioning on a larger θ_{ij} or partial correlation in absolute value results in a higher probability of the edge being present. While such a property clearly holds for the Normal and Laplace spike and slabs, it is not so clear for the MOM spike and slab.

Proposition 14. *Let $\tilde{\pi}^+$ be a regular-SS+ prior with MOM spike and slab. Then $\mathbb{P}_{\tilde{\pi}^+}\left(\gamma_{ij} = 1 \mid \theta_{ij}\right)$ is increasing in $|\theta_{ij}|$ if and only if $\lambda_1 \geq \lambda_0$.*

Let π^+ be a PC-regular-SS+ prior with MOM spike and slab. Then $\mathbb{P}_{\pi^+}\left(\gamma_{ij} = 1 \mid \Delta_{ij}\right)$ is increasing in $|\Delta_{ij}|$ if and only if $\lambda_1 \geq \frac{\lambda_0}{2\lambda_0+1}$.

Proof. Under $\tilde{\pi}^{+}$,

$$\mathbb{P}_{\tilde{\pi}^{+}}\left(\gamma_{ij}=1 \mid \theta_{ij}\right)=\frac{\eta\pi_1\left(\theta_{ij}\right)}{\eta\pi_1\left(\theta_{ij}\right)+(1-\eta)\pi_0\left(\theta_{ij}\right)} \tag{5.7}$$

where

$$\begin{aligned}\pi_1(\theta_{ij}) &= \mathrm{MOM}(\theta_{ij};0,\lambda_1)\\ &= \frac{\theta_{ij}^2}{\lambda_1^{3/2}\sqrt{2\pi}}\exp\left(-\frac{\theta_{ij}^2}{2\lambda_1}\right)\end{aligned}$$

and

$$\begin{aligned}\pi_0(\theta_{ij}) &= \mathrm{N}(\theta_{ij};0,\lambda_0)\\ &= \frac{1}{\lambda_1^{1/2}\sqrt{2\pi}}\exp\left(-\frac{\theta_{ij}^2}{2\lambda_0}\right)\end{aligned}$$

Plugging these into (5.7) and taking the derivative with respect to θ_{ij} gives

$$\frac{\lambda_1^{1/2}\eta(\eta-1)\exp\left(\frac{\theta_{ij}^2(\lambda_0+\lambda_1)}{2\lambda_0\lambda_1}\right)((\lambda_0-\lambda_1)\theta_{ij}^2-2\lambda_0\lambda_1)\theta_{ij}}{\lambda_0^{1/2}\left(\lambda_1^{3/2}(\eta-1)\exp\left(\frac{\theta_{ij}^2}{2\lambda_1}\right)-\lambda_0^{1/2}\eta\theta_{ij}^2\exp\left(\frac{\theta_{ij}^2}{2\lambda_0}\right)\right)}$$

Since the probability (5.7) is continuous in θ_{ij}, $\mathbb{P}_{\tilde{\pi}^{+}}\left(\gamma_{ij}=1 \mid \theta_{ij}\right)$ is increasing in $|\theta_{ij}|$ if and only if this derivative is positive for $\theta_{ij}>0$ and negative for $\theta_{ij}<0$. Notice that the denominator is positive and in the numerator $\lambda_1^{1/2}\eta(\eta-1)\exp\left(\frac{\theta_{ij}^2(\lambda_0+\lambda_1)}{2\lambda_0\lambda_1}\right)$ is negative. Hence we only require that $((\lambda_0-\lambda_1)\theta_{ij}^2-2\lambda_0\lambda_1)\theta_{ij}<0$ when $\theta_{ij}>0$ and $((\lambda_0-\lambda_1)\theta_{ij}^2-2\lambda_0\lambda_1)\theta_{ij}>0$ when $\theta_{ij}<0$, or more simply that $(\lambda_0-\lambda_1)\theta_{ij}^2-2\lambda_0\lambda_1<0$ for all θ_{ij}. This clearly holds if and only if $\lambda_1\geq\lambda_0$.

For the PC-regular-SS+ prior π^{+} we have that

$$\mathbb{P}_{\pi^{+}}\left(\gamma_{ij}=1 \mid \Delta_{ij}\right)=\frac{\frac{\eta}{c_1}\pi_1\left(\Delta_{ij}\right)}{\frac{\eta}{c_1}\pi_1\left(\Delta_{ij}\right)+\frac{(1-\eta)}{c_0}\pi_0\left(\Delta_{ij}\right)}.$$

where $c_i=\int_{-1}^{1}\pi_i(x)\,dx$ are positive constants in Δ_{ij}. Again, $\mathbb{P}_{\pi^{+}}\left(\gamma_{ij}=1 \mid \Delta_{ij}\right)$ is increasing in $|\Delta_{ij}|$ if and only if its derivative is positive when $\Delta_{ij}>0$ and negative when $\Delta_{ij}<0$. The constants c_0, c_1 have little effect on the derivative and the result is a similar condition that $(\lambda_0-\lambda_1)\Delta_{ij}^2-2\lambda_0\lambda_1<0$ for all Δ_{ij}. Noting that $\Delta_{ij}\in(-1,1)$, this holds if and only if $\lambda_1\geq\frac{\lambda_0}{2\lambda_0+1}$

□

Note that the conditions on λ_0, λ_1 in Proposition 14 do not necessarily follow from $\mathrm{Var}_{\pi_1}(\theta_{ij}) > \mathrm{Var}_{\pi_0}(\theta_{ij})$ since $\mathrm{Var}_{\pi_1}(\theta_{ij}) = 3\lambda_1$ and $\mathrm{Var}_{\pi_0}(\theta_{ij}) = \lambda_0$. In the regular-SS+ case the condition $\lambda_1 \geq \lambda_0$ is strictly stronger. In the PC-regular-SS+ case the condition $\lambda_1 \geq \frac{\lambda_0}{2\lambda_0+1}$ is stronger whenever $\lambda_0 < 1$, which would be the case for any sensible choice of spike prior.

5.6 Discussion

In this section we have introduced a general framework for spike and slab priors for Gaussian graphical models. A key benefit of this framework is its flexibility in being able to encode prior beliefs, both on the model space through $\pi(\gamma)$ and on the magnitude of the non-zero partial correlations through π_1. However, this benefit may be lost when the truncation onto the space of positive definite matrices is applied, with the truncation altering the prior marginals on γ and the partial correlations in a way that is tricky to calculate or anticipate. In Propositions 12 and 13 we devised a strategy for removing this negative effect as the problem dimension $p \to \infty$.

An important point for further research is in devising methods for default parameter selection. One approach was proposed by George and McCulloch [1997] based on a threshold of practical relevance. This is a threshold K for which any partial correlation $|\Delta_{ij}| < K$, the modeller would prefer to not include the edge (i, j) in the graphical model, for example in the interest of parsimony. One may include this threshold in parameter setting by allowing, for example, $\pi_1(x) < \pi_0(x) \iff |x| < K$. Further research should involve incorporating the results of Propositions 12 and 13 into such parameter setting.

On the topic of parameter selection, we have generally suggested in this chapter the use of regular-SS+ and PC-regular-SS+ priors with a fixed value for the parameter η, which is set based on prior knowledge of the sparsity of the graphical model. However, in the context of linear regression, Ročková and George [2018] found that the performance of such priors for model selection, particularly when model selection is based on the MAP estimate, is highly sensitive to the specification of η. Instead they suggest a fully Bayesian approach treating η as unknown. Such an extension would fit into the PC-separable-SS+ framework where $\pi(\gamma)$ is allowed to take any form and would be an important consideration for future research.

For PC-regular-SS+ priors we have suggested the diagonal density $\pi_D(\theta_{ii}) \propto \theta_{ii}^{-c}$ for some $c \geq 0$. While this choice leads to scale invariant posterior inference for certain PC-regular-SS+ priors, it should be noted that the resulting prior distribution is improper. However, in a similar way to the Baysian PC-GLASSO as detailed in Section 3.2, there are reasonable assurances that the posterior distribution will be proper provided that $n > 2c$. This is because the $\theta_{ii}^{n/2}$ term in the likelihood combines with θ_{ii}^{-c} so that it no longer goes to infinity as $\theta_{ii} \to 0$. Confirming that the posterior is proper in such a case is another area for future research.

The most important area for further research though, is in the derivation of suitable algorithms for posterior inference. Most simply this might involve using a maximisation algorithm, like the coordinate descent Algorithm 3, to find the MAP estimate under spike and slab priors. However, ultimately a method for posterior sampling of Θ, most likely a Gibbs sampler, will be required to conduct a full posterior analysis and assess the benefits of the different choices for spike and slab prior. Of particular interest is the performance of the MOM spike and slab which has been used to great success in other applications but is novel when applied to Gaussian graphical models.

Chapter 6

Discussion

In this thesis we have demonstrated that essentially all current penalised likelihood methods and prior distributions for Gaussian graphical model selection have a fundamental flaw in that they are not invariant to scalar multiplication of the variables. That is, multiplying the data X by some non-zero diagonal matrix D, for example by changing the unit of measurement for some of the variables, leads to unexpected changes in the estimation of the precision matrix Θ and potentially vastly different graphical model selection. This is a problem because the fundamental conditional independence structure embodied in a graphical model and depicted by it is invariant to such scale transformations.

This problem can be mitigated somewhat by the standardisation of the data, for example by requiring that the data has unit sample variances. The data is invariant to scalar multiplications after standardisation and therefore all methods are also trivially scale invariant. However, data standardisation is rarely suggested, let alone a requirement, in most penalised likelihood and Bayesian methods. It would be quite understandable for an unwitting data scientist without expert knowledge in the area to apply a method such as GLASSO on unstandardised data. In this case the outcome of GLASSO would be highly sensitive to the scale on which variables are measured with variables that happen to have large estimated partial variance likely to have more edges. Furthermore, data standardisation is itself not an innocuous operation. For example, if we believe the non-standardised data to be Gaussian - as we have assumed in many of the examples in this thesis - then the standardised data will not be Gaussian. So these two models are fundamentally different from each other. This is practically as well as methodologically significant. For example we have shown how standardised models perform very poorly for some data generating processes.

A key observation as to why standard methods are not scale invariant is that they are based on the off-diagonal entries of Θ. These off-diagonal entries, θ_{ij}, are themselves not invariant to scalar multiplication and are a poor measure of dependence between the two associated variables X_i, X_j. In particular, θ_{ij} could be large due to a strong dependence between X_i and X_j or because X_i, X_j have small partial variance. Without additional information it is impossible to distinguish these two cases.

From this observation we proposed novel classes of penalty functions and prior distributions that were instead based on partial correlations. Given a suitable penalty or prior on the diagonal entries of Θ, these methods were shown to be scale invariant and therefore return the same graphical model regardless of scale or standardisation. Furthermore, in the case of PC-GLASSO, these methods were shown to perform better in practise when compared to their equivalents based on the off-diagonal entries in both simulated and real data settings.

In each of the previous chapters we suggested potential areas for further research based directly on the content of that chapter and to advance our knowledge of those methods based on partial correlations. To finish this thesis we now present a number of potential areas for future research that are linked to the areas previously discussed, although not directly and with much larger scope.

6.1 Linear models

Although in this thesis we have focused on Gaussian graphical models, it is possible that a similar phenomenon may occur in linear models where the regression coefficients also are not invariant to scalar multiplication. To see this we consider the case which is most similar to a Gaussian graphical model where the response variable Y and covariates X are jointly Gaussian with zero mean and covariance matrix $\Sigma = \Theta^{-1}$. We decompose Σ into

$$\Sigma = \begin{pmatrix} \Sigma_Y & \Sigma_{XY} \\ \Sigma_{XY}^{\mathrm{T}} & \Sigma_X \end{pmatrix}$$

where $\Sigma_Y = \sigma_{11}$, Σ_{XY} is equal to the first row of Σ without σ_{11} and $\Sigma_X = \Theta_X^{-1}$ is equal to the covariance matrix on the X margin. Then the distribution of Y given X is also Gaussian with mean $\Sigma_{XY}\Theta_X X$ and variance $\Sigma_Y - \Sigma_{XY}\Theta_X\Sigma_{XY}^{\mathrm{T}}$. This relates to the linear model

$$Y = X^{\mathrm{T}}\beta + \epsilon$$

where $\beta = \Sigma_{XY}\Theta_X$ and $\epsilon \sim \mathrm{N}(0, \Sigma_Y - \Sigma_{XY}\Theta_X\Sigma_{XY}^{\mathrm{T}})$. Using the fact that $\Sigma\Theta = I$, it can be shown that the entries of β are equal to $\beta_i = -\Sigma_Y \theta_{1i}$. This shows the close relation to Gaussian graphical model selection in this case, where estimation of the regression coefficients is equal to estimation of the first row of Θ, and sparsity in β is the same as sparcity in this first row of Θ.

Direct penalisation of the regression coefficients, as in the LASSO and many other penalised likelihood methods, is therefore analogous to penalisation of the off-diagonal entries of Θ in Gaussian graphical models. The same arguments as have been presented in this thesis can be applied here to show that estimation and model selection under LASSO and similar methods do not satisfy scale invariance under scalar multiplications of the covariates X. By applying the partial correlation framework to such linear models one might instead choose to penalise quantities such as $\theta_{1i}/\sqrt{\theta_{ii}}$.

Of course, this simple example only considers the case where (Y, X) are jointly Gaussian, a special case of the linear model. A significant piece of future work would be to extend such an approach to linear models in general that are robust to scalar multiplication of the covariates. A first step here might be to adapt the LASSO in such a way by applying the L_1 penalty to something other than the regression coefficients directly.

6.2 Positive dependence

Numerous attempts have been made to combine conditional independence relationships with positive (or negative) dependence relationships within a graphical model. Various definitions have been proposed for positive dependence. However the fundamental concept of positive dependence between two variables X_1, X_2 is that observing a 'large' value of X_1 leads to higher probability of X_2 being 'large'.

More formally, some proposed positive dependence definitions are the following: X_1 and X_2 are *positively correlated* if $\mathrm{cov}(X_1, X_2) > 0$; X_1 and X_2 are *associated* if $\mathrm{cov}(f(X_1, X_2), g(X_1, X_2)) \geq 0$ for all non-decreasing functions f and g such that $\mathbb{E}|f(X_1, X_2)|$, $\mathbb{E}|g(X_1, X_2)|$ and $\mathbb{E}|f(X_1, X_2)g(X_1, X_2)|$ all exist; X_1 and X_2 are *multivariate totally positive of order 2* (MTP$_2$) if their joint density function f satisfies $f(x)f(y) \leq f(x \wedge y)f(x \vee y)$ for all x, y, where $x \wedge y$ and $x \vee y$ denote the element-wise minimum and maximum.

Notice that these definitions are all symmetric in the sense that X_1 and X_2 are positively dependent if and only if X_2 and X_1 are positively dependent. One non-symmetric definition is the following: X_1 is stochastically increasing with X_2 if

$F_{X_1|X_2=x_2}(x_1) \leq F_{X_1|X_2=x_2'}(x_1)$ for all x_1 and all $x_2 > x_2'$, where $F_{X_1|X_2=x_2}$ denotes the cdf of X_1 given $X_2 = x_2$.

One early attempt to combine graphical models with positive dependence was in *qualitative probabilistic networks* (QPNs), introduced by Wellman [1990]. Wellman argued that graphical models that assign strictly numeric representations to the distributions of the variables in the model are inappropriately precise for many applications where less strict qualitative constraints may be more realistic. This was achieved by connecting each edge to a sign which determines if the two relevant variables are positively, negatively or otherwise dependent. However, many of the results and dynamics of QPNs in both Wellman [1990] and subsequent papers on QPNs rely on the incorrect assertion that X_1 stochastically increasing with X_2 is equivalent to X_1 and X_2 MTP_2. This is clearly not the case as MTP_2 is a symmetric property but stochastically increasing is not.

A more recent and successful attempt to combine positive dependence and graphical models was by Fallat et al. [2017] who studied graphical models under the assumption that all variables are MTP_2. In a similar vein, Slawski and Hein [2015] and Lauritzen et al. [2017] both investigated the maximum likelihood estimate of a Gaussian precision matrix Θ under the MTP_2 constraint. It was shown that the MTP_2 restriction served as an implicit regulariser and lead to sparsity in the estimate.

A multivariate Gaussian random vector X is MTP_2 if and only if its precision matrix Θ is an M-matrix - i.e. it has positive diagonal entries and non-positive off-diagonals. Equivalently, all partial correlations must be non-negative. This assumption could easily be incorporated into the spike and slab framework of Chapter 5 by simply truncating the slab density on the negative partial correlations $\pi_1(\Delta_{ij})$ between -1 and 0. In this way, the spike density will still represent the (approximately) zero partial correlations, and the slab density will represent the non-zero and positive partial correlations. The MOM density seems a particularly appropriate choice for this.

Further extensions of a similar theme are possible for the spike and slab framework. For example, a slab density that is non-symmetric about zero and with $\mathbb{P}_{\pi_1}(\Delta_{ij} < 0) > \frac{1}{2}$ could be used when a-priori it is expected that most partial correlations are positive. Again, the MOM density is particularly appropriate in this case because the density can simply be re-weighted above and below zero with continuity being maintained. Alternatively, if one wishes to know the sign of the non-zero partial correlations, as well as the graphical model, the spike and slab framework could be adapted to include three levels with $\gamma_{ij} \in \{0, 1, 2\}$. The variable

γ_{ij} now indicates if Δ_{ij} is zero ($\gamma = 0$), positive ($\gamma = 1$) or negative ($\gamma = 2$). The slab can then be spilt into two densities with π_1 truncated on $(-1, 0)$ and π_2 truncated on $(0, 1)$.

6.3 No Simpson's paradox assumption

A necessary condition of MTP$_2$ for a Gaussian random vector is that the covariance matrix Σ has all positive entries. Hence, under the MTP$_2$ condition, all correlations and partial correlations have the same same. MTP$_2$ is, however, a very strong assumption which is often unrealistic especially for large dimension p. In some cases it may be possible to multiply certain variables by -1 to obtain an MTP$_2$ distribution, but in many more cases this is not possible. A strictly weaker assumption that may be of interest is to maintain the condition that all correlations and partial correlations be of the same sign, but relax the condition that all partial correlations are positive. That is, for all i, j, whenever both σ_{ij} and θ_{ij} are non-zero, they are of opposite signs. We call this the *no Simpson's paradox* condition since it means that the sign of dependence is maintained under conditioning on other variables. This assumption is particularly relevant for representing certain types of causal hypotheses and choosing a class of models constrained by these hypotheses. It is often explicitly or implicitly assumed that for a causal relationship to be genuine its directionality would be consistent regardless of conditioning.

The no Simpson's paradox condition corresponds to the set of precision matrices $\Theta = \Sigma^{-1}$ such that for all i, j either $\theta_{ij} = 0$ or $\text{sign}(\theta_{ij}) = -\text{sign}(\sigma_{ij})$. A first step might be to investigate the maximum likelihood estimate under this restriction, as Slawski and Hein [2015] and Lauritzen et al. [2017] did under the MTP$_2$ restriction. However, the space of no Simpson's paradox precision matrices is rather more complicated than MTP$_2$ matrices. A way to approximate (and simplify) the condition is to substitute the covariance matrix Σ by the sample covariance S and restrict Θ to have opposite signs to S, which we call the *sample no Simpson's paradox* condition. In this way the signs of Θ are fixed (given the data) and maximum likelihood estimation under this constraint is more straight forward.

The sample no Simpson's paradox condition actually links to the PC-GLASSO and the convexity of the associated maximisation problem. Recall that in Section 2.9 we noted that the PC-GLASSO penalised likelihood is non-concave, meaning that it does not benefit from the computational advantages of GLASSO. The reason for this non-concavity were terms of the form $-S_{ij}\Delta_{ij}\sqrt{\theta_{ii}\theta_{jj}}$. When S_{ij} and Δ_{ij} are of the same sign, this term is non-concave in θ_{ii} and θ_{jj}. However, making the sample no Simpson's paradox restriction actually results in the penalised likelihood being concave. Hence under this restriction PC-GLASSO could reap the benefits of convex optimisation.

6.4 Conclusion

The project that eventually came to provide the majority of the content for this thesis began with the aim of applying non-local spike and slab priors to Gaussian graphical models. This idea was a direct extension of the spike and slab priors used in Gaussian graphical models by Gan et al. [2018] incorporating the non-local density that has been used in, for example Avalos-Pacheco et al. [2020]. Although a modest aim, the vision for this project was to eventually extend the model to incorporate more nuanced prior beliefs, such as positive dependence, and to incorporate ideas of causality. However, when beginning this project we came to find that the use of spike and slab priors for Gaussian graphical models contains some unique issues which, to our knowledge, have not so far been addressed in the literature.

The first of these issues was the truncation onto the space of positive definite matrices that is required when placing a prior on the precision matrix Θ. As demonstrated in Chapter 5, this truncation particularly effects spike and slab priors due to its impact on the prior edge inclusion probability potentially leading to the prior placing higher probability on more sparse graphs than intended. This issue is also unique to the Gaussian graphical model setting since other settings, for example linear regression, do not require such a truncation of the prior distribution. The identification of this issue led to the theoretical results based on Wigner matrices in Section 5.3.

The second issue was discovered when applying the BAGUS method of Gan et al. [2018] to data generated using a precision matrix Θ with non-unit diagonal in the star graph setting. We found that the estimates provided by BAGUS and the performance of its model selection were highly dependent on the diagonal entries of Θ - i.e. the method is not scale invariant. This led to the decision to place the spike and slab priors on the partial correlations rather than directly to Θ.

We eventually found that this issue of non-scale invariance was actually a more general issue common to essentially all penalised likelihood methods and prior distributions used for Gaussian graphical model selection. In particular, non-scale invariance effects the well known and popular GLASSO method - something we later found is much maligned by academics familiar with Gaussian graphical models but which is not generally discussed in the literature. This naturally gave the idea of adapting the GLASSO in the same way as the spike and slab priors by placing the L_1 penalty on the partial correlations. After discovering that this idea, to the best of our knowledge, has not already been investigated, this quickly became the focus of the thesis.

The result has been classes of partial correlation based penalised likelihoods and prior distributions which are invariant to scalar multiplication of the variables - a property which is fundamental to conditional independence and graphical models. The PC-GLASSO is one specific example from this class of penalised likelihoods which is a directly analogous to GLASSO but achieves scale invariance and from our applications is shown to offer improvements over GLASSO. We hope that this thesis have served to demonstrate the benefits of partial correlation based methods for Gaussian graphical model selection and precision matrix estimation and that this will lead to further research and improvements in their use.

Appendix A

Maximisation problem in Algorithm 2

Step 2 of Algorithm 2, requires the maximisation of (2.5) with respect to $\Delta_{ij}, \theta_{ii}, \theta_{jj}$ whilst all other variables are held fixed. In this appendix we give details of how this maximum may be found as well as demonstrating that the updating a positive definite matrix as in Step 2 of Algorithm 2 retains positive definiteness. To ease notation let $x = \Delta_{ij}$, $y_1 = \sqrt{\theta_{ii}}$ and $y_2 = \sqrt{\theta_{jj}}$. The objective function is

$$\begin{aligned} f(x, y_1, y_2) = {} & \log(ax^2 + bx + c) + 2c_n(\log(y_1) + \log(y_2)) \\ & - y_1^2 - y_2^2 - 2c_{12}xy_1y_2 - 2c_1y_1 - 2c_2y_2 - 2\rho|x|, \end{aligned}$$

where

$$\begin{aligned} c_n &= 1 - \frac{4}{n}, \\ c_{12} &= S_{ij}, \\ c_1 &= \sum_{k \neq i,j} S_{ik}\Delta_{ik}\sqrt{\theta_{kk}}, \\ c_2 &= \sum_{k \neq i,j} S_{jk}\Delta_{jk}\sqrt{\theta_{kk}}. \end{aligned}$$

The $\log(ax^2 + bx + c)$ term comes from the $\log\det(\Delta)$, since the determinant of a symmetric matrix is quadratic in the off-diagonal entries. The coefficients (a, b, c) do not have a simple closed-form, as they depend on the matrix determinant, but they can be easily obtained by evaluating the determinant of Δ for three different values of Δ_{ij} (faster methods for computing these determinants are possible since they only involve changing a single entry) and solving the resulting system of equations. The range of values that x is can take given by

$$(l,u) := \{x : ax^2 + bx + c > 0\} \cap (-1,1).$$

Any value of x in this set ensures positive definiteness of Δ. This is because Δ is positive definite if and only if all its leading principal minors are positive. WLOG, letting Δ_{ij} be in the bottom row of Δ, if the previous estimate is positive definite then the first $p-1$ leading principal minors are positive. The condition $ax^2 + bx + c > 0$ ensures that the final leading principal minor, $\det(\Delta)$, is also positive. The maximisation problem can then be expressed as

$$\begin{aligned} \max_{x,y_1,y_2} \quad & f(x,y_1,y_2) \\ \text{s.t.} \quad & x \in (l,u) \\ & y_1, y_2 > 0 \end{aligned} \tag{A.1}$$

We denote the partial derivatives of f by

$$f_x(x,y_1,y_2) = \frac{2ax+b}{ax^2+bx+c} - 2c_{12}y_1y_2 - 2\rho\text{sign}(x), \quad x \neq 0,$$

$$f_{y_1}(x,y_1,y_2) = 2c_n y_1^{-1} - 2y_1 - 2c_{12}xy_2 - 2c_1,$$

$$f_{y_2}(x,y_1,y_2) = 2c_n y_2^{-1} - 2y_2 - 2c_{12}xy_1 - 2c_2,$$

To solve this problem we consider separately the cases $c > 0$ and $c \leq 0$.

Case $c > 0$.

We begin by looking at the case $c > 0$, which implies that $0 \in (l,u)$. We split the problem into three sections, finding local maxima in $x = 0$, $x \in (0,u)$, $x \in (l,0)$ separately and then selecting from these the global maximum.

Optimization for $x = 0$.

Let $x = 0$. By setting $f_{y_1}(x,y_1,y_2) = 0$ and $f_{y_2}(x,y_1,y_2) = 0$ we get that the optimal values of (y_1,y_2) are

$$y_1 = \frac{1}{2}\left(\sqrt{c_1^2 + 4c_n} - c_1\right),$$

$$y_2 = \frac{1}{2}\left(\sqrt{c_2^2 + 4c_n} - c_2\right).$$

Optimization over $x > 0$.

Let $x \in (0, u)$. Setting $f_{y_1}(x, y_1, y_2) = 0$ gives

$$x = \frac{c_n y_1^{-1} - y_1 - c_1}{c_{12} y_2}, \tag{A.2}$$

and setting $f_{y_2}(x, y_1, y_2) = 0$ along with (A.2) gives

$$y_2 = \frac{1}{2}\left(-c_2 \pm \sqrt{c_2^2 + 4(y_1^2 + c_1 y_1)}\right). \tag{A.3}$$

Using (A.2)-(A.3) one can write $f_x(x, y_1, y_2)$ in terms of only y_1 and solve $f_x(x, y_1, y_2) = 0$ numerically to obtain the stationary points. The range of y_1 values to search in the numerical solving of $f_x(x, y_1, y_2) = 0$ can be found by considering the constraints $x \in (0, u)$, $y_1, y_2 > 0$ as well as (A.2) and (A.3).

The constraint $x < u$ results in some condition on the following quartic which we refer to as $q(y_1)$

$$\left(1 - \frac{1}{u^2 c_{12}^2}\right) y_1^4 + \left(c_1 - \frac{2c_1}{u^2 c_{12}^2} + \frac{c_2}{u c_{12}}\right) y_1^3 + \left(\frac{2c_n}{u^2 c_{12}^2} - \frac{c_1^2}{u^2 c_{12}^2} + \frac{c_1 c_2}{u c_{12}}\right) y_1^2 \\ + \left(\frac{2c_1 c_n}{u^2 c_{12}^2} - \frac{c_2 c_n}{u c_{12}}\right) y_1 - \frac{c_n^2}{u^2 c_{12}^2} \tag{A.4}$$

We first summarize the range of y_1 values that needs to be considered, depending on the values of (c_{12}, c_2), and subsequently outline their derivation. If the positive root is taken in (A.3) for y_2 then the following constraints are required

1. $y_1 < \frac{1}{2}\left(-c_1 + \sqrt{c_1^2 + 4c_n}\right)$, if $c_{12} > 0$
2. $y_1 > \frac{1}{2}\left(-c_1 + \sqrt{c_1^2 + 4c_n}\right)$, if $c_{12} < 0$
3. $y_1 \geq \frac{1}{2}\left(-c_1 + \sqrt{c_1^2 - c_2^2}\right)$ or $y_1 \leq \frac{1}{2}\left(-c_1 - \sqrt{c_1^2 - c_2^2}\right)$
4. $y_1 > -c_1$, if $c_2 > 0$
5. If $c_{12} > 0$, either $y_1 > \frac{1}{2}\left(\frac{1}{2}uc_{12}c_2 - c_1 + \sqrt{\left(c_1 - \frac{1}{2}uc_{12}c_2\right)^2 + 4c_n}\right)$ or $q(y_1) > 0$
6. If $c_{12} < 0$, either $y_1 < \frac{1}{2}\left(\frac{1}{2}uc_{12}c_2 - c_1 + \sqrt{\left(c_1 - \frac{1}{2}uc_{12}c_2\right)^2 + 4c_n}\right)$ or $q(y_1) < 0$

The negative root in (A.3) must only be considered if $c_2 < 0$ and $y_1 < -c_1$ (also implying that $c_1 < 0$ and, from constraint 1, $c_{12} > 0$). In this case the inequalities in constraints 5 and 6 must be reversed.

We outline how to obtain the above constraints. The constraint $x > 0$ along with (A.2) implies that

$$\text{sign}(y_1^2 + c_1 y_1 - c_n) = -\text{sign}(c_{12}).$$

Hence, if $c_{12} > 0$ then the range of values to consider can be restricted to

$$y_1 < \frac{1}{2}\left(-c_1 + \sqrt{c_1^2 + 4c_n}\right),$$

giving constraint 1, while if $c_{12} < 0$ then the inequality is reversed giving constraint 2. Note that if $c_{12} = 0$ then the optimisation problem is simpler and so the details of this case are omitted.

For y_2 to take a real value in (A.3) we must have $4y_1^2 + 4c_1 y_1 + c_2^2 \geq 0$ which implies that either

$$y_1 \geq \frac{1}{2}\left(\sqrt{c_1^2 - c_2^2} - c_1\right),$$

or

$$y_1 \leq \frac{1}{2}\left(-\sqrt{c_1^2 - c_2^2} - c_1\right).$$

giving constraint 3.

Combining the constraint $y_2 > 0$ with (A.3), if $c_2 > 0$ then we need $y_1 \geq -c_1$ in order for there to be a solution for y_2, giving constraint 4. On the other hand, if $c_2 < 0$ and $0 < y_1 < -c_1$ then there are two solutions for y_2 and one must consider both the positive and negative roots in (A.3). For all other situations one must only consider the positive root.

Now combining the constraint $x < u$ with (A.2) and (A.3), one obtains the inequality

$$\frac{2}{uc_{12}}\left(c_n y_1^{-1} - y_1 - c_1\right) + c_2 < \sqrt{c_2^2 + 4(y_1^2 + c_1 y_1)}$$

from which constraints 5 and 6 follow.

Combining each of these constraints give the range of possible values for y_1 to numerically search for a stationary point. Once y_1 is found, (A.3) and (A.2) give the corresponding (x, y_2). Note that it is possible that there be no stationary points within $x > 0$.

Optimization over $x < 0$.

Finding stationary points in the interval $x \in (l, 0)$ is analogous to the case where $x \in (0, u)$, but with some sign changes and so the details are omitted.

Case $c \leq 0$.

Consider the case where $c \leq 0$. Then it is easy to see that when $b > 0$ then $(l, u) \subseteq (0, 1)$, while if $b < 0$ then $(l, u) \subseteq (-1, 0)$. Again, solving this is very similar to the previous case, however one must pay closer attention to the range of values y_1 may take. In particular, when $b > 0$, (A.2) must still hold at stationary points, but one must restrict this in (l, u) rather than $(0, u)$. This results in two quartic constraints on y_1. Again the details are omitted.

Bibliography

Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. An Introduction to Random Matrices. *An Introduction to Random Matrices*, 2009. doi: 10.1017/cbo9780511801334.

Alejandra Avalos Pacheco. *Factor regression for dimensionality reduction and data integration techniques with applications to cancer data*. PhD thesis, University ofWarwick, 2018.

Alejandra Avalos-Pacheco, David Rossell, and Richard S. Savage. Heterogeneous Large Datasets Integration Using Bayesian Factor Regression. *Bayesian Analysis*, -1(-1):1–34, 2020. ISSN 1931-6690. doi: 10.1214/20-ba1240.

Jonathan J. Azose and Adrian E. Raftery. Estimating large correlation matrices for international migration. *The Annals of Applied Statistics*, 12(2):940–970, 2018. doi: 10.1214/18-AOAS1175.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre D'Aspremont. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *Journal of Machine Learning Research*, 9:485–516, 2008. ISSN 02552930.

Sayantan Banerjee and Subhashis Ghosal. Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:147–162, 2015. ISSN 10957243. doi: 10.1016/j.jmva.2015.01.015. URL `http://dx.doi.org/10.1016/j.jmva.2015.01.015`.

Haim Y. Bar, James G. Booth, and Martin T. Wells. A Scalable Empirical Bayes Approach to Variable Selection in Generalized Linear Models. *Journal of Computational and Graphical Statistics*, 29(3):535–546, 2020. ISSN 15372715. doi: 10.1080/10618600.2019.1706542.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Leo Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24(6):2350–2383, 1996. ISSN 00905364. doi: 10.1214/aos/1032181158.

Peter Bühlmann and Lukas Meier. Discussion: One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1534–1541, 2008. ISSN 00905364. doi: 10.1214/07-AOS0316A.

Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011. ISSN 01621459. doi: 10.1198/jasa.2011.tm10155.

Alexandre Calon, Elisa Espinet, Sergio Palomo-Ponce, Daniele V F Tauriello, Mar Iglesias, María Virtudes Céspedes, Marta Sevillano, Cristina Nadal, Peter Jung, Xiang H-F Zhang, Daniel Byrom, Antoni Riera, David Rossell, Ramón Mangues, Joan Massague, Elena Sancho, Eduard Batlle, and Or Elena Sancho. Dependency of colorectal cancer on a TGF-beta-driven programme in stromal cells for metastasis initiation. *Cancer Cell*, 22(5):571–584, 2012. doi: 10.1016/j.ccr.2012.08.013.Dependency. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3512565/pdf/nihms-422793.pdf`.

Emmanuel J. Candès, Michael B. Wakin, and Stephen P. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14 (5-6):877–905, 2008. ISSN 10695869. doi: 10.1007/s00041-008-9045-x.

Jack Storror Carter, David Rossell, and Jim Q. Smith. Partial Correlation Graphical LASSO. pages 1–41, 2021. URL `http://arxiv.org/abs/2104.10099`.

Ismaël Castillo and Aad Van Der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Annals of Statistics*, 40(4):2069–2101, 2012. ISSN 00905364. doi: 10.1214/12-AOS1029.

Davide Chicco and Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1):1–13, 2020. ISSN 14712164. doi: 10.1186/s12864-019-6413-7.

Guido Consonni, Dimitris Fouskakis, Brunero Liseo, and Ioannis Ntzoufras. Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13(2):627–679, 2018. ISSN 19316690. doi: 10.1214/18-BA1103.

R. Dennis Cook and Liliana Forzani. On the mean and variance of the generalized inverse of a singular wishart matrix. *Electronic Journal of Statistics*, 5:146–158, 2011. ISSN 19357524. doi: 10.1214/11-EJS602.

A. P. Dawid and S. L. Lauritzen. Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics*, 21(3):1272–1317, 1993.

Lee Dicker, Baosheng Huang, and Xihong Lin. Variable selection and estimation with the seamless-L0 penalty. *Statistica Sinica*, 23(2):929–962, 2013. ISSN 10170405. doi: 10.5705/ss.2011.074.

Adrian Dobra, Alex Lenkoski, and Abel Rodriguez. Bayesian inference for general gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433, 2011. ISSN 01621459. doi: 10.1198/jasa.2011.tm10465. URL `https://doi.org/10.1198/jasa.2011.tm10465`.

Shaun Fallat, Steffen Lauritzen, Kayvan Sadeghi, Caroline Uhler, Nanny Wermuth, and Piotr Zwiernik. *Total positivity in Markov structures*, volume 45. 2017. ISBN 2011300975. doi: 10.1214/16-AOS1478.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001. ISSN 1537274X. doi: 10.1198/016214501753382273.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics*, 3(2):521–541, 2009. ISSN 19326157. doi: 10.1214/08-AOAS215.

Peter G.M. Forbes and Steffen Lauritzen. Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra and Its Applications*, 473:261–283, 2015. ISSN 00243795. doi: 10.1016/j.laa.2014.08.015. URL `http://dx.doi.org/10.1016/j.laa.2014.08.015`.

Rina Foygel and Mathias Drton. Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*, pages 1–9, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008. ISSN 14654644. doi: 10.1093/biostatistics/kxm045.

Lingrui Gan, Naveen N Narisetty, and Feng Liang. Bayesian Regularization for Graphical Models With Unequal Shrinkage. *Journal of the American Statistical Association*, 114(527):1218–1231, 2018. ISSN 1537274X. doi: 10.1080/01621459.2018.1482755.

Xin Gao, Daniel Q. Pu, Yuehua Wu, and Hong Xu. Tuning parameter selection for penalized likelihood estimation of inverse covariance matrix. *Statistica Sinica*, 22: 1123–1146, 2012. URL `http://arxiv.org/abs/0909.0934`.

Edward I. George and Robert E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. ISSN 1537274X. doi: 10.1080/01621459.1993.10476353.

Edward I. George and Robert E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997. ISSN 10170405.

Paolo Giudici and Peter J. Green. Decomposable graphical Gaussian model determination. *Biometrika*, 86(4):785–801, 1999. ISSN 00063444. doi: 10.1093/biomet/86.4.785.

Min Jin Ha and Wei Sun. Partial Correlation Matrix Estimation using Ridge Penalty Followed by Thresholding and Reestimation. *Biometrics*, 70(3):762–770, 2014. doi: 10.1111/biom.12186.

Søren Højsgaard, David Edwards, and Steffen Lauritzen. *Graphical Models with R*. Springer Science & Business Media, 2012.

Valen E. Johnson and David Rossell. Non-Local Prior Densities for Default Bayesian Hypothesis Tests. *Journal of the Royal Statistical Society B*, 72:143–170, 2010.

Valen E. Johnson and David Rossell. Bayesian Model Selection in High-Dimensional Settings Valen. *Journal of the American Statistical Association*, 107(498):649–660, 2012. doi: 10.1080/01621459.2012.682536.Bayesian.

Zakaria S. Khondker, Hongtu Zhu, Haitao Chu, Weili Lin, and Joseph G. Ibrahim. The Bayesian covariance lasso. *Statistics and its Interface*, 6:243–259, 2013.

Tõnu Kollo and Kaire Ruul. Approximations to the distribution of the sample correlation matrix. *Journal of Multivariate Analysis*, 85(2):318–334, 2003. ISSN 0047259X. doi: 10.1016/S0047-259X(02)00037-4.

Steffen Lauritzen, Caroline Uhler, and Piotr Zwiernik. Maximum likelihood estimation in Gaussian models under total positivity. 2017. URL `http://arxiv.org/abs/1702.04031`.

Steffen L. Lauritzen. *Graphical models.* Clarendon Press, Oxford, 1996. ISBN 0-190852219-3.

Yunfan Li, Bruce A. Craig, and Anindya Bhadra. The Graphical Horseshoe Estimator for Inverse Covariance Matrices. *Journal of Computational and Graphical Statistics*, 28(3):747–757, 2019. ISSN 15372715. doi: 10.1080/10618600.2019.1575744. URL `https://doi.org/10.1080/10618600.2019.1575744`.

Heng Lian. Shrinkage tuning parameter selection in precision matrices estimation. *Journal of Statistical Planning and Inference*, 141(8):2839–2848, 2011. ISSN 03783758. doi: 10.1016/j.jspi.2011.03.008. URL `http://dx.doi.org/10.1016/j.jspi.2011.03.008`.

Lina Lin, Mathias Drton, and Ali Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10(1): 806–854, 2016. ISSN 19357524. doi: 10.1214/16-EJS1126.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of Graphical Models.* CRC Press, Boca Raton, 1st edition, 2018. ISBN 9780429463976. doi: https://0-doi-org.pugwash.lib.warwick.ac.uk/10.1201/9780429463976.

David Madigan, Jeremy York, and Denis Allard. Bayesian Graphical Models for Discrete Data. *International Statistical Review / Revue Internationale de Statistique*, 63(2):215, 1995. ISSN 03067734. doi: 10.2307/1403615.

Benjamin M. Marlin and Kevin P. Murphy. Sparse Gaussian graphical models with unknown block structure. *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, pages 705–712, 2009.

Benjamin M. Marlin, Mark Schmidt, and Kevin P. Murphy. Group sparse priors for covariance estimation. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, pages 383–392, 2009.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006. ISSN 00905364. doi: 10.1214/009053606000000281.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. ISSN 1537274X. doi: 10.1080/01621459.1988.10478694.

A. Mohammadi and E. C. Wit. Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015. ISSN 19316690. doi: 10.1214/14-BA889.

Andrei Patrascu and Ion Necoara. Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization. *Journal of Global Optimization*, 61(1):19–46, 2015. ISSN 15732916. doi: 10.1007/s10898-014-0151-9.

Judea Pearl. *Causality*. Cambridge university press, 2009. ISBN 0-521-77362-8.

Judea Pearl and Azaria Paz. Graphoids: graph-based logic for reasoning about relevance relations. *Advances in Artificial Intelligence*, 2:357–363, 1987.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009. ISSN 01621459. doi: 10.1198/jasa.2009.0126.

Pradeep Ravikumar, Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Model selection in Gaussian graphical models: High-dimensional consistency of l 1-regularized MLE. *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, pages 1329–1336, 2009.

Veronika Ročková and Edward I. George. The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444, 2018. ISSN 1537274X. doi: 10.1080/01621459.2016.1260469. URL https://doi.org/10.1080/01621459.2016.1260469.

David Rossell and Piotr Zwiernik. Dependence in elliptical partial correlation graphs. 2020. URL http://arxiv.org/abs/2004.13779.

Fabian Scheipl, Ludwig Fahrmeir, and Thomas Kneib. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532, 2012. ISSN 01621459. doi: 10.1080/01621459.2012.737742.

Guiling Shi, Chae Young Lim, and Tapabrata Maiti. Model selection using mass-nonlocal prior. *Statistics and Probability Letters*, 147: 36–44, 2019. ISSN 01677152. doi: 10.1016/j.spl.2018.11.027. URL `https://doi.org/10.1016/j.spl.2018.11.027`.

Martin Slawski and Matthias Hein. Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and Its Applications*, 473:145–179, 2015. ISSN 00243795. doi: 10.1016/j.laa.2014.04.020.

Milan Studeny. *Probabilistic conditional independence structures.* Springer Science & Business Media, 2006. ISBN 1-85233-891-1.

Matyas a. Sustik and Ben Calderhead. GLASSOFAST: An efficient GLASSO implementation. Technical report, 2012.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso, 1996. ISSN 00293970. URL `https://statweb.stanford.edu/~tibs/lasso/lasso.pdf`.

Ivan Vujačić, Antonino Abbruzzo, and Ernst Wit. A computationally fast alternative to cross-validation in penalized Gaussian graphical models. *Journal of Statistical Computation and Simulation*, 85(18):3628–3640, 2015. ISSN 15635163. doi: 10.1080/00949655.2014.992020. URL `https://doi.org/10.1080/00949655.2014.992020`.

Hao Wang. Bayesian graphical lasso models and eficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012. ISSN 19360975. doi: 10.1214/12-BA729.

Hao Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015. ISSN 19316690. doi: 10.1214/14-BA916.

Lingxiao Wang, Xiang Ren, and Quanquan Gu. Precision matrix estimation in high dimensional gaussian graphical models with faster rates. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, 51:177, 2016.

Michael P. Wellman. Fundamental Concepts of Qualitative Probabilistic Networks. *Artificial Intelligence*, 44(3):257–303, 1990.

Joe Whittaker. *Graphical Models in Applied Multivariate Statistics.* Wiley, 1990. ISBN 978-0-471-91750-2.

Donald R Williams. Beyond Lasso: A Survey of Nonconvex Regularization in Gaussian Graphical Models. Technical report, 2020.

By Frederick Wong and Christopher K Carter. Efficient Estimation of Covariance Selection Models Author (s): Frederick Wong , Christopher K . Carter and Robert Kohn Published by : Oxford University Press on behalf of Biometrika Trust Stable URL : https://www.jstor.org/stable/30042090 REFERENCES Li. 90(4):809–830, 2003.

Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015. ISSN 14364646. doi: 10.1007/s10107-015-0892-3. URL `http://dx.doi.org/10.1007/s10107-015-0892-3`.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. ISSN 00063444. doi: 10.1093/biomet/asm018.

Cun Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, 2010. ISSN 00905364. doi: 10.1214/09-AOS729.

Shuheng Zhou, Sara van de Geer, and Peter Bühlmann. Adaptive Lasso for High Dimensional Regression and Gaussian Graphical Modeling. 2009. URL `http://arxiv.org/abs/0903.2515`.

Hui Zou and Runze Li. One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Annals of Statistics*, 36(4):1509–1533, 2008. doi: 10.1214/009053607000000802.