Received: 28 May 2021 | Revised: 29 May 2023 | Accepted: 31 May 2023

DOI: 10.1111/sjos.12675

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Partial correlation graphical LASSO

Jack Storror Carter[1,2] | **David Rossell**[3,4] | **Jim Q. Smith**[2,5]

[1]Department of Mathematics, University of Genova, Genova, Italy

[2]Department of Statistics, University of Warwick, Coventry, UK

[3]Department of Business and Economics, Universitat Pompeu Fabra, Barcelona, Spain

[4]Data Science Center, Barcelona Graduate School of Economics, Barcelona, Spain

[5]The Alan Turing Institute, London, UK

Correspondence
Jack Storror Carter, Department of Mathematics, University of Genova, Genova, Italy.
jack.carter@dima.unige.it

Funding information
Agencia Estatal de Investigación, Grant/Award Number: CNS2022-135963; Bando Per L'incentivazione Della Progettazione Europea 2020, Grant/Award Number: 100021-2020-Er-Incent_Eu_Riccomagno; Engineering and Physical Sciences Research Council, Grant/Award Numbers: EP/K039628/1, EP/L016710/1, EP/N510129/1; Europa Excelencia, Grant/Award Number: EUR2020-112096; Fundación BBVA, Ayudas a Proyectos de Investigación Científica en Matemáticas 2021

Abstract

Standard likelihood penalties to learn Gaussian graphical models are based on regularizing the off-diagonal entries of the precision matrix. Such methods, and their Bayesian counterparts, are not invariant to scalar multiplication of the variables, unless one standardizes the observed data to unit sample variances. We show that such standardization can have a strong effect on inference and introduce a new family of penalties based on partial correlations. We show that the latter, as well as the maximum likelihood, L_0 and logarithmic penalties are scale invariant. We illustrate the use of one such penalty, the partial correlation graphical LASSO, which sets an L_1 penalty on partial correlations. The associated optimization problem is no longer convex, but is conditionally convex. We show via simulated examples and in two real datasets that, besides being scale invariant, there can be important gains in terms of inference.

KEYWORDS

covariance matrix estimation, Gaussian graphical model, graphical LASSO, partial correlation, penalized likelihood, precision matrix

1 INTRODUCTION

In Gaussian graphical models, most popular frequentist approaches to sparse estimation of the precision matrix penalize the absolute value of the entries of the precision matrix. Gaussian graphical models are invariant to scalar multiplication of the variables; however, it is well-known that such penalization approaches do not share this property. We show that the only scale-invariant strategies, within a large class of precision matrix penalties, are the logarithmic and L_0 penalties. It is possible to address this issue via a data preprocessing step of standardizing the data to have unit sample variances. However, as we illustrate next, this standardization can adversely affect inference. In this paper, we propose a family of methods based on partial correlations and show that they ensure scale invariance without requiring this standardization step.

As motivation we present a simple example where the goal is to estimate the entries in a $p \times p$ precision matrix Θ. We set $p = 50$ and generate $n = 100$ independent Gaussian draws with zero mean and covariance Θ^{-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. This is a setting in which recovering the graphical model is relatively straightforward, for example in Yuan and Lin (2007). The top left panel in Figure 1 shows the regularization path for the estimated partial correlations when applying GLASSO (Friedman et al., 2008) to the unstandardized data. For a large range of values for the regularization parameter ρ the truly zero θ_{ij}'s are completely separated from the nonzeroes. However, the top right panel shows that when standardizing the data to unit sample variances

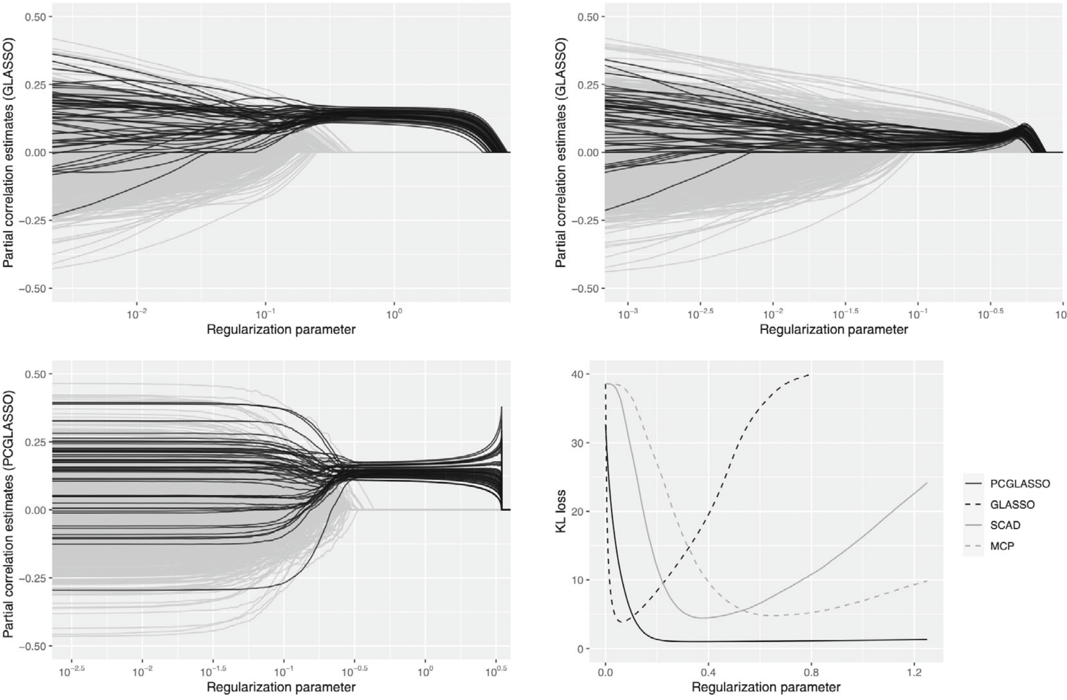

FIGURE 1 Top: partial correlation regularization paths for GLASSO in the $p = 50$ star graph example on the original data (left), and standardized data (right). Estimates of truly nonzero θ_{ij} are in black. Bottom: Partial correlation regularization paths for PCGLASSO in the $p = 50$ star graph example (left) and Kullback–Leibler (KL) loss over the regularization paths for different penalties applied to standardized data (right).

the quality of the inference suffers. In particular the true graphical model is not recovered for any ρ. The bottom left panel shows the results obtained by applying a LASSO penalty to the partial correlations, our proposed PCGLASSO, which as we show is scale invariant. The bottom right panel demonstrates how the estimation accuracy measured by Kullback–Leibler (KL) loss (see Section 8) of GLASSO and two other methods reviewed below suffer in comparison to PCGLASSO when using standardized data.

Lack of invariance is not restricted to the GLASSO, but, as we show later, affects essentially all continuous penalties on the precision matrix, as well as standard prior distributions in Bayesian settings.

The paper is organized as follows. Section 2 sets notation and reviews popular classes of likelihood penalties which we refer to as *regular* penalty functions, and their Bayesian equivalents, *regular* prior distributions. Section 3 introduces a class of penalties and prior distributions on partial correlations, and the PCGLASSO as a particular case. Section 4 shows that the PCGLASSO, as well as the logarithmic and L_0 penalties are scale invariant, while other regular penalty functions are not. Section 5 offers an alternative argument for standardizing the data when using regular penalties, related to situations where the likelihood function is exchangeable in two partial correlations, hence one may wish for inference to be exchangeable as well. Section 6 compares the related prior distributions of GLASSO and PCGLASSO and Section 7 discusses computational issues for the PCGLASSO and gives a certain conditional convexity result. Section 8 shows examples on simulated, gene expression and stock market datasets. We end the paper with a short discussion.

2 | PENALIZED LIKELIHOOD IN GAUSSIAN GRAPHICAL MODELS

Let $X = (X^{(1)}, \dots, X^{(p)}) \sim \mathrm{N}(\mu, \Sigma)$ 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 $(X_1, \dots, X_n)$ of X and denote their sample covariance by S (note that this is the biased sample covairance with denominator n). 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 maximize a penalized likelihood function of the form $l(\Theta|S) - Pen(\Theta)$, where

$$l(\Theta|S) = \frac{n}{2}\left[\log(\det(\Theta)) - \mathrm{tr}(S\Theta) - p\log(2\pi)\right], \tag{1}$$

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 1. A penalty function $\mathrm{Pen}(\Theta)$ is *separable* if

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

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

A separable penalty is *regular* if $\text{pen}_{ii} = \text{pen}_{jj}$ 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

$$\Theta^{\rho}_{\text{GLASSO}}(S) = \arg\max \log(\det(\Theta)) - \text{tr}(S\Theta) - \rho \sum_{i=1}^{p} \sum_{j=1}^{p} |\theta_{ij}| \tag{2}$$

for some given regularization parameter $\rho \geq 0$. Meinshausen and Bühlmann (2006) proposed 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 parameterizing (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 smoothly clipped absolute deviation (SCAD) penalty (Fan et al., 2009; Fan & Li, 2001) and the minimax concave penalty (MCP) (Wang et al., 2016; Zhang, 2010), 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 $\text{Pen}(\Theta) = \rho \sum_{i<j} \mathbb{I}(\theta_{ij} \neq 0)$.

The adaptive LASSO (Fan et al., 2009; Zhou et al., 2009) is an important example of a non-regular penalty. It uses an L_1 penalty where weights depend on the data via some initial estimate of Θ, and hence does not satisfy Definition 1. However, as noted by Bühlmann et al. (2008) and Candes et al. (2008), the adaptive LASSO can be seen as a first-order approximation of the logarithmic penalty where $\text{pen}_{ij}(\theta_{ij}) = \rho \log(|\theta_{ij}|)$, which is regular. Both papers propose an iterative version of adaptive LASSO that formally targets this logarithmic penalty.

There is a well known equivalence between penalized likelihood and maximum a posteriori estimates in Bayesian frameworks. In particular, the estimate under a penalty (Pen) is equal to the mode of the posterior distribution under the prior density $\pi(\Theta) \propto \exp(-\text{Pen}(\Theta))\mathbb{I}(\Theta \in \mathcal{S})$ where $\mathcal{S}$ is the set of symmetric, positive definite matrices. With this in mind we define *separable* and *regular prior distributions*.

Definition 2. A prior distribution with density π on Θ is *separable* if

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

where π_{ii} is a density function with support $(0, \infty)$ and π_{ij} is a density function with support $\mathbb{R}$ which are nonincreasing 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.

The correspondence between penalized likelihoods and prior distributions has been utilized by the Bayesian LASSO regression of Park and Casella (2008) and Hans (2009) and in Gaussian graphical models by Wang (2012) and Khondker et al. (2013). Of particular interest to this paper, Wang (2012) showed that under the GLASSO prior the marginal prior distribution of partial correlations does not depend on the regularization parameter. We explore this further in Section 6. The Bayesian interpretation has also been used to create new penalties functions, for example, by Banerjee and Ghosal (2015) and Gan et al. (2019), both of whom set mixture priors on the entries of Θ.

3 | PARTIAL CORRELATION GRAPHICAL LASSO

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

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

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 $\theta = \operatorname{diag}(\Theta)$ and Δ is the matrix with unit diagonal and off-diagonal entries Δ_{ij}. The penalized likelihood function then becomes

$$\frac{n}{2}\left[\log(\det(\Delta)) + \sum_i \log(\theta_{ii}) - \operatorname{tr}(S\theta^{\frac{1}{2}}\Delta\theta^{\frac{1}{2}})\right] - \operatorname{Pen}(\theta, \Delta). \tag{3}$$

We believe that partial correlations are a better measure of dependence than the precision matrix entries 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 parameterization, a corresponding prior class, and subsequently state our PCGLASSO as a particular case.

Definition 3. A penalty (Pen) is *partial correlation separable* (PC-separable) if it is of the form

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

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

A PC-separable penalty function is *symmetric* if $\operatorname{pen}_{ii} = \operatorname{pen}_{jj}$ 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 3 includes formulations that do not penalize the diagonal entries, that is, $\operatorname{pen}_{ii}(\theta_{ii}) = 0$. Note also that the L_0 and logarithmic penalties are PC-separable since $\theta_{ij} = 0$ if and only if $\Delta_{ij} = 0$ and $\log(|\theta_{ij}|) = \log(|\Delta_{ij}|) + \frac{1}{2}\log(\theta_{ii}) + \frac{1}{2}\log(\theta_{jj})$.

Definition 4. A prior $\pi(\theta, \Delta)$ is *(symmetric) PC-separable* if the penalty function $\operatorname{Pen}(\theta, \Delta) = -\log(\pi(\theta, \Delta))$ is (symmetric) PC-separable.

Any PC-separable prior can be written as

$$\pi(\theta, \Delta) \propto \prod_i \pi_{ii}(\theta_{ii}) \prod_{i<j} \pi_{ij}(\Delta_{ij}) \mathbb{I}(\Delta \in \mathcal{S}_1), \tag{4}$$

where $\mathcal{S}_1$ is the set of symmetric, positive definite matrices with unit diagonal.

PCGLASSO is a symmetric PC-separable penalty applying the L_1 norm to the partial correlations $\operatorname{pen}_{ij}(\Delta_{ij}) = n\rho|\Delta_{ij}|$, and a logarithmic penalty to the diagonal $\operatorname{pen}_{ii}(\theta_{ii}) = 2\log(\theta_{ii})$. The penalized likelihood function, after removing constants, is given by

$$\log(\det(\Delta)) + \left(1 - \frac{4}{n}\right)\sum_i \log(\theta_{ii}) - \operatorname{tr}\left(S\theta^{\frac{1}{2}}\Delta\theta^{\frac{1}{2}}\right) - \rho\sum_{i\neq j}|\Delta_{ij}|. \tag{5}$$

The logarithmic penalty on the diagonal entries ensures scale invariance of the PCGLASSO (Section 4). A coefficient of 2 is used since in the univariate $p = 1$ case this minimizes the mean squared error of the estimated precision amongst logarithmic penalties (see Appendix A). Although many methods use the same penalty forms for diagonal and off-diagonal entries, it seems natural to use different forms since the former do not aim to induce sparsity. For example, Yuan and Lin (2007) argued for a GLASSO framework where one does not penalize the diagonal.

As is common for LASSO-type penalties, one may choose a sequence of regularization parameters ρ for which to calculate the PCGLASSO estimate $\hat{\Theta}(\rho)$ and select the solution that maximizes some suitable criterion. In Section 8 we used a Bayesian information like criterion (BIC), which selects the estimate minimizing

$$\mathrm{BIC}(\hat{\Theta}(\rho), S) = \log(n) \sum_{i<j} \mathbb{I}(\hat{\theta}_{ij}(\rho) \neq 0) - 2\, l(\hat{\Theta}(\rho)|S). \tag{6}$$

Parameter selection via such a BIC is common within penalized likelihoods and has been used in Gaussian graphical models by, for example, Yuan and Lin (2007) and Lian (2011). It has also been shown to provide consistent graphical model selection when used with the SCAD penalty (Gao et al., 2012). Other potential criteria that have been explored for GLASSO are cross validation and the extended Bayesian information criterion (EBIC, Foygel & Drton, 2010), which we also consider in our real data applications. Further discussion is available in, for example, Vujačić et al. (2015).

There are some examples of penalty functions for Gaussian graphical models based on partial correlations. Ha and Sun (2014) utilized a ridge penalty. The space method of Peng et al. (2009), similarly to PCGLASSO, 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 allowing one to specify prior beliefs on partial correlations.

4 | 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 nonzero diagonal, then DX is also Gaussian with precision matrix

$$\Theta_D = D^{-1}\Theta D^{-1}. \tag{7}$$

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 (7) 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 PC-separable penalties more generally are. Recall that any estimator can be made scale invariant by standardizing the data to unit sample variances prior to obtaining the estimate, but as discussed this has an effect on inference. We start by defining two notions of scale invariance related to the point estimate and to the recovered graphical structure.

Definition 5. An estimator $\hat{\Theta}$ is *scale invariant* if for any sample covariance matrix S and any diagonal $p \times p$ matrix D with nonzero diagonal entries,

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

$\hat{\Theta}$ 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 (7). 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. Note that the results could equivalently be written in terms of the maximum a posteriori estimate under corresponding prior distributions. All proofs are in Appendix B.

Proposition 1. *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.*

In particular, the GLASSO, SCAD, and MCP estimators are not scale invariant. Further, as illustrated in Figure 1 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 1.

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 shows the associated GLASSO estimates $\Theta^{\rho}_{\text{GLASSO}}(S)$. The right panel considers the situation where the data were 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).

As shown in Proposition 1, 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 PCGLASSO is scale invariant.

Proposition 2. *Any estimator based on a symmetric PC-separable penalty is scale invariant, provided* $\text{pen}_{ii}(\theta_{ii}) = c\log(|\theta_{ii}|)$ *for some constant $c \geq 0$.*

In the Bayesian framework, Proposition 2 implies scale invariance of the a posteriori mode under symmetric PC-separable priors. That is, let $\tilde{\Theta} = \hat{\Theta}(DSD)$ be 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 $\pi\left(\tilde{\Theta}|DSD\right)$ and $\pi\left(D\tilde{\Theta}D|S\right)$.

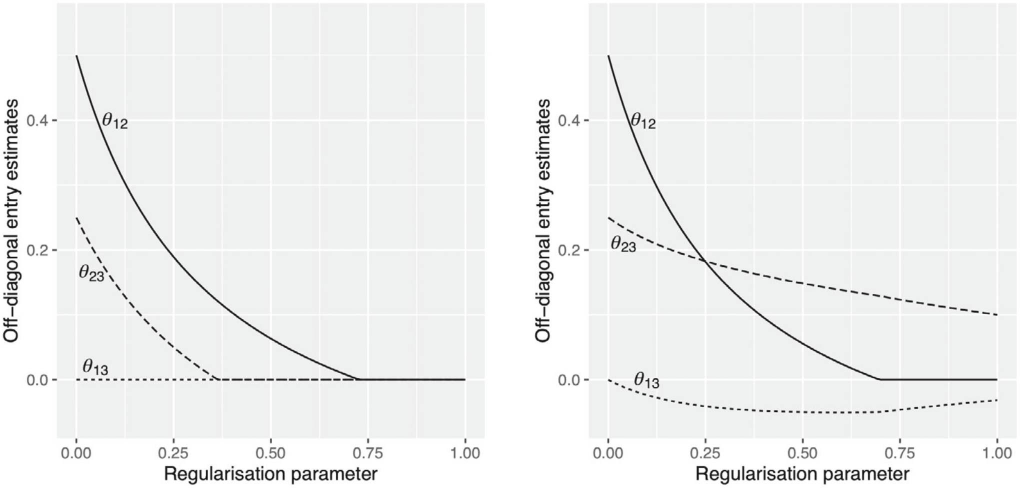

FIGURE 2 Estimated off-diagonal entries $\Theta^{\rho}_{\text{GLASSO}}(S)$ (left) and $D\Theta^{\rho}_{\text{GLASSO}}(DSD)D$ (right) for regularization parameter $\rho \in [0, 1]$.

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 6. Let $\pi(\Theta)$ be a prior density, S a sample covariance and D a diagonal matrix with nonzero diagonal. Let the posterior density associated to S be $\pi(\Theta|S) \propto L(\Theta|S)\pi(\Theta)$, and that associated to DSD be $\pi(\Theta|DSD) \propto L(\Theta|DSD)\pi(\Theta)$ where L is the Gaussian likelihood function.

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

$$\mathbb{P}_{\pi}(\Theta \in A|DSD) = \mathbb{P}_{\pi}(\Theta \in A_D|S), \tag{8}$$

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

In particular, (8) implies that the two posterior distributions on the partial correlations Δ are equal up to appropriate sign changes, that is, when D has all positive entries, $\pi(\Delta|S) = \pi(\Delta|DSD)$ (since Δ associated to Θ is equal to that associated to $D\Theta D$).

Proposition 3. *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.*

We note that the proof of Proposition 3 does not depend on $\pi_{ij}(\Delta_{ij})$ being nonincreasing or π_{ij} being the same for all $i \neq j$. Hence the result extends to any prior of the form (4) 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—that is, 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.

5 | EXCHANGEABLE INFERENCE

We now discuss an alternative view on the desirability of standardizing the data when using regular penalties, based on notions of exchangeable inference. The simplest situation occurs when the likelihood function is exchangeable in two or more Δ_{ij}'s, for example when two rows in the sample correlation matrix $R = \mathrm{diag}(S)^{-1/2} S \mathrm{diag}(S)^{-1/2}$ 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 Example 1, regular penalties can lead to significantly different inference (unless one standardizes the data).

Example 1. 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 likelihood is also exchangeable in the remaining Δ_{ij} and their estimates should ideally be close to 0, their true value.

The left panel of Figure 3 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 standardizes the data, so that the sample covariance is equal to R, one obtains the center panel of Figure 3. Now $\hat{\Delta}_{12} = \hat{\Delta}_{13} = \hat{\Delta}_{14}$ for any regularization 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 PCGLASSO estimates (on either the original or standardized data, due to scale invariance) in the right panel of Figure 3 satisfy $\hat{\Delta}_{12} = \hat{\Delta}_{13} = \hat{\Delta}_{14}$, and the truly zero parameters are clearly distinguished.

We remark that the notion can be extended to conditional exchangeability, that is, the likelihood being symmetric in $(\Delta_{ij}, \Delta_{kl})$ given the remaining parameters in Δ and θ. For example, the likelihood is conditionally exchangeable in $(\Delta_{ij}, \Delta_{ik})$ when the sample covariances and precisions are related by the same constant, that is, $S_{ij} = cS_{ik}$ and $\theta_{kk}^{1/2} = c\theta_{jj}^{1/2}$ for some $c > 0$, and the partial correlations with other variables are equal, that is, $\Delta_{jl} = \Delta_{kl}$ for all $l \notin \{i, j, k\}$. See Appendix E for additional details and supplementary results. Conditional exchangeability would be relevant in situations where two variables (j, k) have the same estimated partial correlations

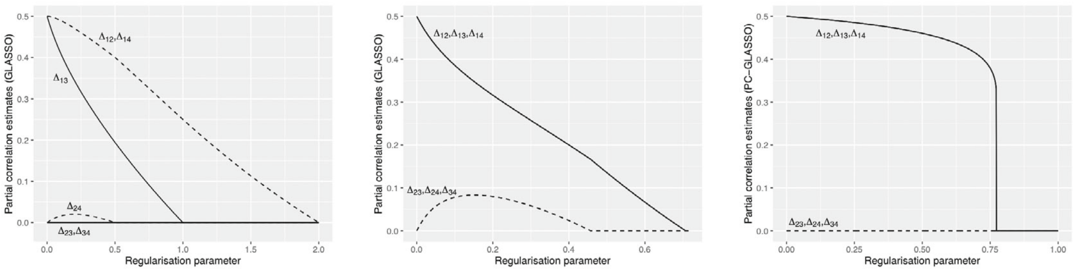

FIGURE 3 Partial correlation regularization paths in $p = 4$ star graph example for GLASSO on the original S (left), standardized S (center) and PCGLASSO (right).

with all other variables (e.g., zero), as well as the same sample covariances with a third variable i. In such situations, one may wish for equal inference, in particular equal point estimates $\hat{\Delta}_{ij} = \hat{\Delta}_{ik}$.

6 | GLASSO AND PCGLASSO PRIOR DISTRIBUTIONS

In this section we provide further insights into the shrinkage induced by GLASSO and PCGLASSO, by comparing their implied prior distributions in a Bayesian framework.

The GLASSO prior (Wang, 2012) can be written as

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

where $\lambda = n\rho$, whereas the PCGLASSO prior is given by

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

To illustrate the effect of increasing the parameter λ for fixed $p = 5$ (Wang (2012) provides results for growing p with fixed λ), we sampled from each prior via rejection sampling for $\lambda = 1, 2$, and 4. Figure 4 plots the marginal densities of Δ_{12} and θ_{11}. The top left panel verifies the claim of Wang (2012) that the GLASSO prior on partial correlations $\pi_G(\Delta_{ij})$ does not depend on λ, whereas the bottom panel shows that $\pi_G(\theta_{ii})$ is shrunk toward 0 as λ increases. In contrast, the PCGLASSO 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 regularization of the partial correlations, while GLASSO does so by shrinking the diagonal θ_{ii}.

7 | COMPUTATION

An important feature of GLASSO is its defining of a convex problem that significantly facilitates computation and its theoretical study. For example, Friedman et al. (2008) related GLASSO to a

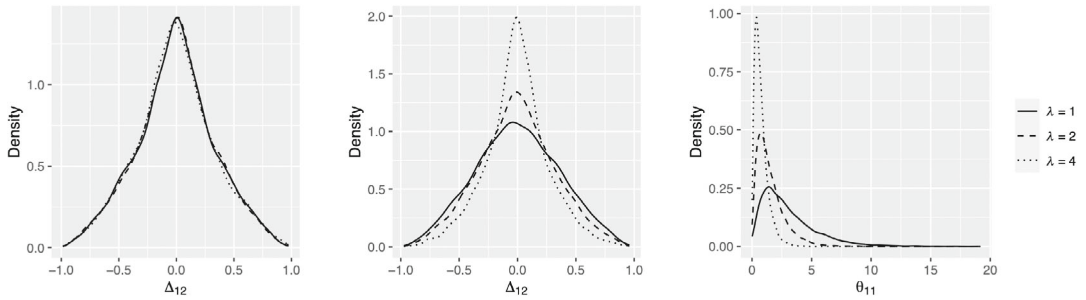

FIGURE 4 Marginal prior densities for the partial correlations under GLASSO prior (left) and PCGLASSO prior (centre) and for the diagonal entries under the GLASSO prior (right).

sequence of LASSO problems, and Sustik and Calderhead (2012) provided improved algorithms. Computation for nonconvex penalties such as SCAD and MCP poses a harder challenge, but the Local Linear Approximation of Zou (2008) and Fan et al. (2009) greatly facilitates this task. The PCGLASSO optimization problem is nonconvex, however, it is conditionally convex given $\theta = \mathrm{diag}(\Theta)$.

Proposition 4. *The penalized likelihood function (5) is concave in Δ, for any fixed value of θ.*

Proposition 4 (proof in Appendix C) opens the possibility to consider block optimization algorithms of the form described in Algorithm 1, where $\hat{\theta}$ and $\hat{\Delta}$ are updated sequentially. This takes advantage of the conditional convexity of the problem, allowing for convex optimization methods to be exploited in Step 3 of Algorithm 1. In fact, the optimization in Step 3 is very similar to the GLASSO optimization problem, however, it is not completely analogous due to Δ having fixed unit diagonal.

In our examples, we took an even simpler strategy and used a coordinate descent algorithm. Despite its conceptual simplicity, the algorithm requires careful updating of each parameter to ensure positive definiteness of $\hat{\Delta}$. For brevity we defer details to Appendix D. For the scale of problems addressed in this paper, provided the starting point is close to the optimum, the algorithm typically converges in a few iterations.

Algorithm 1. Blockwise optimization algorithm

1. Choose start points $\Delta^{(0)}, \theta^{(0)}$.
2. Let $\theta^{(1)}$ maximize

$$f(\Delta, \theta) = \log(\det(\Delta)) + \left(1 - \frac{4}{n}\right) \sum_i \log(\theta_{ii}) - \mathrm{tr}\left(S\theta^{1/2}\Delta\theta^{1/2}\right) - \rho\|\Delta\|_1,$$

   for fixed $\Delta = \Delta^{(0)}$.
3. Let $\Delta^{(1)}$ maximize $f(\Delta, \theta)$ for fixed $\theta = \theta^{(1)}$.
4. Update $\Delta^{(0)} = \Delta^{(1)}, \theta^{(0)} = \theta^{(1)}$.
5. Repeat Steps 2–4 until some stopping condition is achieved and then output $\Delta = \Delta^{(1)}, \theta = \theta^{(1)}$.

8 | APPLICATIONS

We now assess the performance of PCGLASSO against GLASSO, SCAD, and MCP, setting the regularization parameters via the BIC in (6). For GLASSO we use the version with no penalization on the diagonal entries as this generally has improved performance. 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 we standardized data to unit sample variances, and rescaled the estimates via (7). GLASSO was implemented using the R package **glasso** and SCAD and MCP using the package **GGMncv** (Williams, 2020).

Our primary interest is studying PCGLASSO versus GLASSO, as they are directly comparable in the sense of using the same L_1 penalty structure. We consider SCAD and MCP as benchmarks designed to ameliorate the estimation bias associated to the L_1 penalty. Although not considered here for brevity, it would also be interesting to study the use of SCAD and MCP penalties on partial correlations.

8.1 | Simulations

When choosing a simulation setting for Gaussian graphical models, it can be challenging to tune the data generating Θ to achieve a desired difficulty of the inference problem. The fact that methods, such as the PCGLASSO, exist that are invariant to scalar multiplications suggests that the difficulty of estimating Θ and detecting non-zero θ_{ij} is also invariant to such rescalings. In the θ, Δ parameterization, this would mean that the problem's difficulty depends only on Δ and not on θ. Furthermore, it suggests that it is the magnitude of the partial correlations Δ_{ij} rather than the off-diagonal θ_{ij} that drive the difficulty of the inference. Further empirical evidence for this is provided in Appendix F where we consider two simulation exercises. First is a simple $p = 2$ dimensional setting where the diagonals θ_{ii} grow but the partial correlation stays constant, and we observe that the mean squared error of the maximum likelihood estimator (MLE) remains constant. The second exercise is a star graph simulation where we observe that model selection performance deteriorates as the nonzero partial correlations decrease; however, the model selection remains constant for varying diagonal entries when the partial correlations are fixed. This supports the claim that it is the magnitude of the partial correlations rather than of the partial variances or covariances that drive the problem's hardness. Of course there are additional factors that effect the problem difficulty such as the graphical model structure (if it is decomposable and, if so, the clique sizes) and the eigenvalues of Δ which is an area for further investigation.

We considered four simulation scenarios with Gaussian data, truly zero mean and precision matrix Θ with unit diagonal (so that the off-diagonals are equal to the partial correlations) and off-diagonal entries as follows.

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

Scenario 2: Hub graph - Partition variables into four 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: 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: 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 setting we used $p = 20$ variables, considered sample sizes $n \in \{30, 100\}$ and we performed 100 independent simulations. We also investigated $p = 50$ variables with $n = 100$ samples performing 50 independent simulations. To assess estimation accuracy we used the KL loss

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

and the Frobenius norm (F-norm)

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

To assess model selection accuracy we considered the Matthews correlation coefficient (MCC)

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

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 can be found in, for example, Chicco and Jurman (2020). We report the mean of these metrics over the independent simulations as well as the SE.

The results of the simulations are summarized in Tables 1–4, also including sensitivity and specificity. PCGLASSO generally outperformed GLASSO in all scenarios, and either outperformed or was competitive to SCAD and MCP. More specifically, PCGLASSO strongly outperformed other methods in the Star graph setting in estimation and model selection. The Star graph is an example where there is a large range in the node degrees, suggesting that penalizing partial correlations can be particularly beneficial in such situations. The AR2 model is the opposite situation where every node has either one or two edges. Here PCGLASSO still improved significantly over GLASSO, and to a lesser extent over SCAD or MCP in the $n = 30$ case, but for $n = 100$ the latter two provided better estimation and model selection recovery. PCGLASSO was also generally better in the Hub and Random graph settings, particularly for $n = 30$, although SCAD and MCP offered slight improvements for $n = 100$. The $p = 50$ case demonstrates that these results also hold in higher dimensions.

Figure 5 shows the proportion of the 100 simulations in which each edge was selected, illustrating that PCGLASSO generally selected sparser models than GLASSO, particularly in the Star and Hub scenarios.

TABLE 1 Star results.

$p = 20, n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
PCGLASSO	1.42 (0.35)	1.69 (0.58)	0.978 (0.043)	0.999 (0.008)	0.995 (0.010)
GLASSO	2.59 (0.36)	3.36 (1.42)	0.270 (0.044)	0.866 (0.105)	0.578 (0.057)
SCAD	8.07 (3.78)	10.87 (4.76)	0.344 (0.136)	0.738 (0.143)	0.764 (0.079)
MCP	8.58 (4.11)	11.60 (5.17)	0.335 (0.126)	0.737 (0.138)	0.756 (0.079)
$p = 20, n = 100$	**FNorm**	**KL**	**MCC**	**Sensitivity**	**Specificity**
PCGLASSO	0.70 (0.11)	0.46 (0.12)	0.993 (0.017)	1 (0)	0.999 (0.004)
GLASSO	1.66 (0.09)	1.20 (0.13)	0.304 (0.019)	0.996 (0.014)	0.508 (0.031)
SCAD	1.33 (0.38)	1.01 (0.38)	0.739 (0.135)	0.958 (0.046)	0.926 (0.049)
MCP	1.39 (0.40)	1.09 (0.41)	0.737 (0.128)	0.952 (0.050)	0.928 (0.043)
$p = 50, n = 100$	**FNorm**	**KL**	**MCC**	**Sensitivity**	**Specificity**
PCGLASSO	1.07 (0.11)	1.10 (0.21)	0.998 (0.006)	1 (0)	1 (0)
GLASSO	2.65 (0.08)	3.66 (0.19)	0.240 (0.015)	0.911 (0.033)	0.674 (0.011)
SCAD	3.01 (0.43)	4.68 (0.81)	0.529 (0.071)	0.851 (0.051)	0.935 (0.017)
MCP	3.26 (0.46)	5.17 (0.88)	0.515 (0.070)	0.843 (0.052)	0.932 (0.018)

Abbreviations: KL, Kullback–Leibler; MCC, Matthews correlation coefficient.

TABLE 2 Hub results.

$p = 20, n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
PCGLASSO	1.85 (0.29)	2.83 (0.74)	0.696 (0.081)	0.988 (0.043)	0.917 (0.034)
GLASSO	2.26 (0.21)	3.11 (0.64)	0.469 (0.071)	0.998 (0.012)	0.763 (0.066)
SCAD	7.80 (4.43)	11.55 (6.33)	0.339 (0.110)	0.830 (0.108)	0.715 (0.115)
MCP	8.22 (4.68)	12.30 (6.64)	0.329 (0.111)	0.821 (0.112)	0.707 (0.125)
$p = 20, n = 100$	**FNorm**	**KL**	**MCC**	**Sensitivity**	**Specificity**
PCGLASSO	0.91 (0.15)	0.70 (0.20)	0.858 (0.069)	1 (0)	0.969 (0.019)
GLASSO	1.63 (0.19)	1.11 (0.22)	0.483 (0.054)	1 (0)	0.778 (0.048)
SCAD	0.91 (0.21)	0.55 (0.20)	0.918 (0.062)	0.998 (0.012)	0.984 (0.014)
MCP	0.91 (0.22)	0.55 (0.22)	0.920 (0.066)	0.997 (0.014)	0.984 (0.015)
$p = 50, n = 100$	**FNorm**	**KL**	**MCC**	**Sensitivity**	**Specificity**
PCGLASSO	1.56 (0.14)	2.15 (0.39)	0.796 (0.053)	1 (0)	0.981 (0.007)
GLASSO	3.34 (0.15)	4.48 (0.47)	0.508 (0.037)	1 (0)	0.913 (0.015)
SCAD	1.92 (0.24)	2.62 (0.60)	0.724 (0.074)	0.989 (0.018)	0.971 (0.012)
MCP	1.93 (0.26)	2.76 (0.68)	0.693 (0.069)	0.983 (0.019)	0.966 (0.013)

Abbreviations: KL, Kullback–Leibler; MCC, Matthews correlation coefficient.

TABLE 3 AR2 results.

$p = 20, n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
PCGLASSO	3.64 (0.31)	5.26 (0.62)	0.283 (0.093)	0.301 (0.194)	0.922 (0.077)
GLASSO	3.83 (0.15)	5.77 (0.55)	0.219 (0.128)	0.126 (0.140)	0.984 (0.032)
SCAD	5.98 (4.47)	9.17 (5.56)	0.290 (0.105)	0.444 (0.162)	0.837 (0.114)
MCP	6.09 (4.61)	9.48 (5.87)	0.270 (0.105)	0.432 (0.159)	0.832 (0.110)
$p = 20, n = 100$	**FNorm**	**KL**	**MCC**	**Sensitivity**	**Specificity**
PCGLASSO	2.30 (0.33)	2.00 (0.38)	0.530 (0.052)	0.855 (0.094)	0.774 (0.069)
GLASSO	2.72 (0.36)	2.22 (0.49)	0.520 (0.057)	0.818 (0.122)	0.785 (0.092)
SCAD	1.60 (0.23)	1.33 (0.29)	0.767 (0.065)	0.908 (0.059)	0.918 (0.039)
MCP	1.60 (0.23)	1.37 (0.31)	0.785 (0.065)	0.895 (0.062)	0.932 (0.035)
$p = 50, n = 100$	**FNorm**	**KL**	**MCC**	**Sensitivity**	**Specificity**
PCGLASSO	4.78 (0.29)	7.80 (0.82)	0.502 (0.029)	0.673 (0.071)	0.926 (0.024)
GLASSO	5.47 (0.15)	9.04 (0.69)	0.535 (0.033)	0.598 (0.059)	0.956 (0.017)
SCAD	3.30 (0.29)	5.50 (0.73)	0.664 (0.049)	0.790 (0.067)	0.955 (0.016)
MCP	3.35 (0.29)	5.78 (0.73)	0.650 (0.050)	0.760 (0.061)	0.957 (0.016)

Abbreviations: KL, Kullback–Leibler; MCC, Matthews correlation coefficient.

TABLE 4 Random graph results.

$p = 20, n = 30$	FNorm	KL	MCC	Sensitivity	Specificity
PCGLASSO	2.30 (0.25)	3.07 (0.51)	0.336 (0.091)	0.310 (0.153)	0.951 (0.041)
GLASSO	2.38 (0.18)	3.49 (0.61)	0.311 (0.118)	0.205 (0.158)	0.978 (0.040)
SCAD	4.87 (4.31)	6.56 (4.81)	0.206 (0.094)	0.318 (0.113)	0.876 (0.078)
MCP	5.12 (3.83)	6.98 (4.47)	0.194 (0.092)	0.320 (0.112)	0.868 (0.078)
$p = 20, n = 100$	**FNorm**	**KL**	**MCC**	**Sensitivity**	**Specificity**
PCGLASSO	1.43 (0.16)	1.23 (0.25)	0.572 (0.059)	0.614 (0.110)	0.941 (0.029)
GLASSO	1.62 (0.15)	1.37 (0.27)	0.581 (0.061)	0.641 (0.102)	0.941 (0.030)
SCAD	1.32 (0.15)	1.08 (0.23)	0.598 (0.070)	0.610 (0.105)	0.952 (0.029)
MCP	1.32 (0.14)	1.09 (0.22)	0.594 (0.070)	0.587 (0.110)	0.957 (0.027)
$p = 50, n = 100$	**FNorm**	**KL**	**MCC**	**Sensitivity**	**Specificity**
PCGLASSO	2.53 (0.16)	4.52 (0.49)	0.536 (0.038)	0.713 (0.048)	0.944 (0.014)
GLASSO	3.45 (0.19)	5.43 (0.69)	0.588 (0.048)	0.823 (0.044)	0.936 (0.020)
SCAD	2.79 (0.20)	5.23 (0.54)	0.509 (0.048)	0.627 (0.059)	0.952 (0.017)
MCP	2.89 (0.21)	5.62 (0.57)	0.471 (0.050)	0.594 (0.064)	0.948 (0.019)

Abbreviations: KL, Kullback–Leibler; MCC, Matthews correlation coefficient.

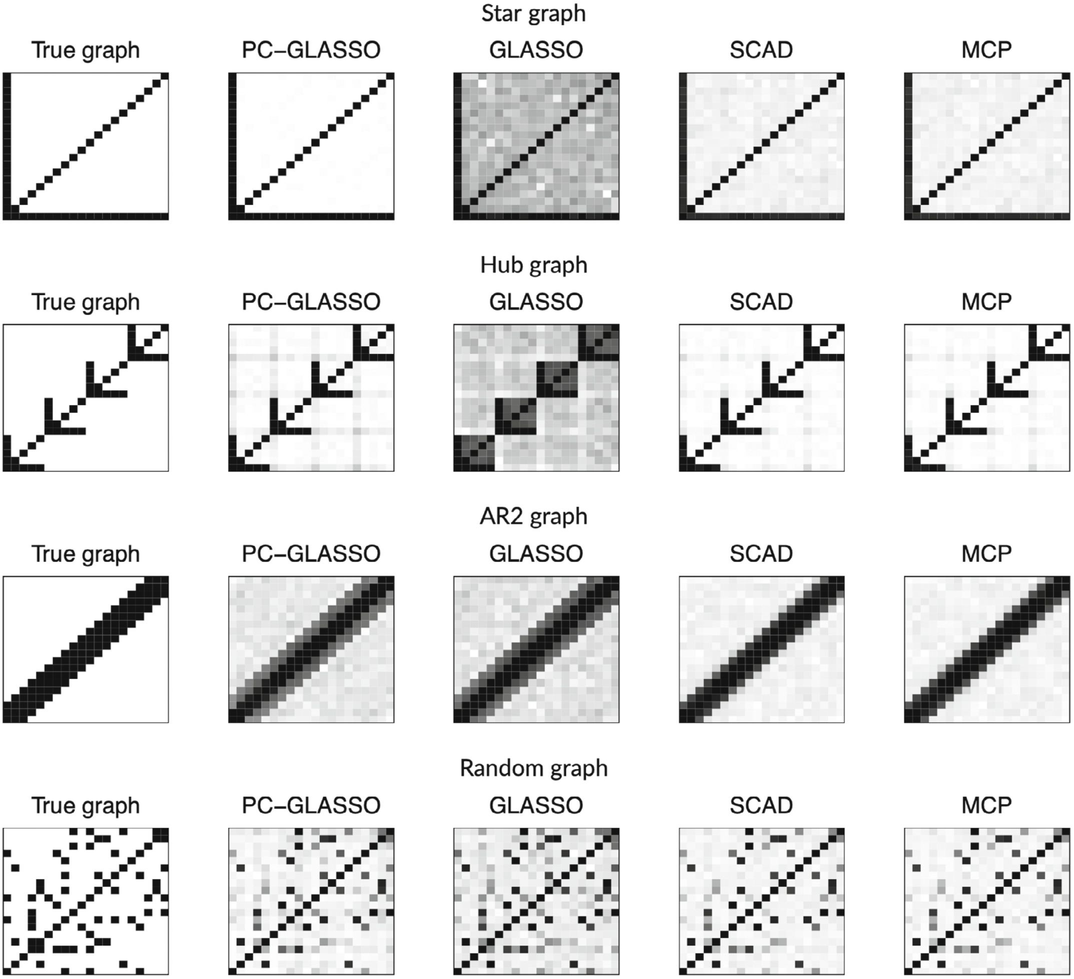

FIGURE 5 Proportion of simulations in which each edge was selected.

8.2 | Gene expression data

We assessed the predictive performance of the four penalized likelihood methods in the gene expression data of Calon et al. (2012). The data contain 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.

Figure 6 (left) plots the model size versus test sample log-likelihood, and indicates the models chosen by the BIC and EBIC. For both these solutions, PCGLASSO achieved a significantly higher log-likelihood than the other three methods, and selected a model of roughly comparable size.

8.3 | Stock market data

We analyzed the stock market data in the R package **huge**, investigated in the graphical model context by Banerjee and Ghosal (2015). The data contain daily closing stock prices of companies

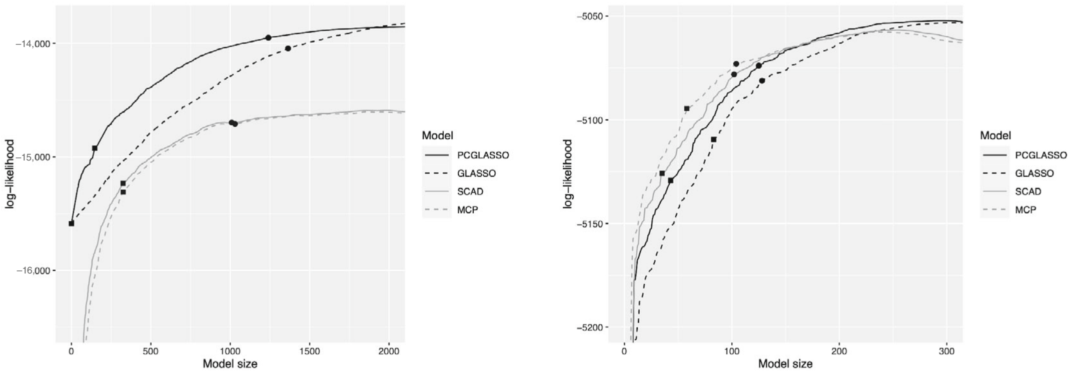

FIGURE 6 Model size versus predictive ability in the gene expression (left) and stock market (right) data. Estimates selected via Bayesian information like criterion and extended Bayesian information criterion with $\gamma = 0.5$ are shown by dots and squares, respectively.

in the S&P 500 index between January 1, 2003 and January 1, 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, $\tilde{X}_{jt} = \log\left(Y_{j,t+1}/Y_{jt}\right)$ the log-returns, and $X_{jt} = \tilde{X}_{jt} - \overline{X}_t$ the de-trended returns, where $\overline{X}_t = \sum_{j=1}^{p} \tilde{X}_{jt}$. 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 five-sample SDs away from the mean in any of the p variables. There remained 1121 observations of which we randomly selected 1000 for the training and 121 for the test data.

Figure 6 (right) shows the results, which highlight interesting trade-offs in sparsity versus predictive accuracy. PCGLASSO selected a smaller model than GLASSO for BIC and EBIC, and achieved a higher log-likelihood in the test data for any model with <200 edges, whereas GLASSO attained a higher log-likelihood at the selected model. Interestingly, the SCAD and MCP penalties provided a similar accuracy to PCGLASSO, albeit slightly higher for models with <150 edges and slightly lower for larger models.

9 | DISCUSSION

Penalized 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 showed that notions of exchangeability also motivate the need for standardizing the data when using regular penalties.

Standardizing the data is not innocuous. First, even when the variables follow a Gaussian distribution, that is no longer the case for the scaled variables, which exhibit thicker tails. Second, as demonstrated in several of our examples, applying regular penalties to scaled data can adversely affect inference. This effect was particularly detrimental in examples where the true underlying graph has a large range in node degrees, as in the Star graph setting.

A wide class of PC-separable penalties, including the PCGLASSO, overcome these issues as they are scale invariant and do not require standardization. Using a Bayesian viewpoint, we illustrated that PCGLASSO induces a different shrinkage than standard penalties, in that the former induces shrinkage on partial correlations, whereas the latter do not. Our examples showed that such differential shrinkage can offer significant improvements both in estimation and model selection.

A limitation of our work lies in the computation. While the efficiency of the coordinate descent algorithm is reasonable in lower dimensions, the computations become impractical for larger p. However, the conditional convexity of the PCGLASSO problem opens interesting strategies for future improvements.

Further interesting future work is to investigate the theoretical properties of PCGLASSO, for example model selection consistency, which holds for GLASSO only under certain nontrivial conditions (Raskutti et al., 2008). The wider set of PC-separable penalties also warrant further exploration, most obviously PC-separable versions of the SCAD and MCP penalties. On the Bayesian side, a PC-separable version of the spike and slab penalty of Gan et al. (2019) may also be of interest. Beyond the Gaussian case, penalization of partial correlations also seems natural for partial correlation graphs in elliptical and transelliptical distributions (Rossell & Zwiernik, 2021).

ACKNOWLEDGMENTS

Jack Storror Carter was funded by the EPSRC grant EP/L016710/1 as part of the Oxford-Warwick Statistics Programme (OxWaSP) and is now funded by the University of Genova under the 100021-BIPE 2020 grant. David Rossell was partially funded by Ayudas Fundación BBVA Proyectos de Investigación Científica en Matemáticas 2021, grant Consolidación investigadora CNS2022-135963 by the AEI, and Europa Excelencia EUR2020-112096 from the AEI/10.13039/501100011033 and European Union "NextGenerationEU"/PRT. Jim Q. Smith was supported by the Alan Turing Institute EPSRC Grant EP/N510129/1 and and by the EPSRC grant EP/K039628/1.

ORCID

Jack Storror Carter https://orcid.org/0000-0002-2766-7732

REFERENCES

Azose, J. J., & Raftery, A. E. (2018). Estimating large correlation matrices for international migration. *The Annals of Applied Statistics*, *12*, 940.

Banerjee, O., El Ghaoui, L., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, *9*, 485–516.

Banerjee, S., & Ghosal, S. (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, *136*, 147–162.

Bühlmann, P., Meier, L., & Zou, H. (2008). Discussion of "one-step sparse estimates in nonconcave penalized likelihood model" by H. Zou and R. Li. *Annals of Statistics*, *36*, 1534–1541.

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D. V., Iglesias, M., Céspedes, M. V., Sevillano, M., Nadal, C., Jung, P., Zhang, X. H.-F., Byrom, D., Riera, A., Rossell, D., Mangues, R., Massagué, J., Sancho, E., & Batlle, E. (2012). Dependency of colorectal cancer on a tgf-β-driven program in stromal cells for metastasis initiation. *Cancer Cell*, *22*, 571–584.

Candes, E. J., Wakin, M. B., & Boyd, S. P. (2008). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, *14*, 877–905.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 1–13.

Fan, J., Feng, Y., & Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, *3*, 521–541.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Foygel, R., & Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. *Advances in Neural Information Processing Systems*, *23*.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*, 432–441.

Gan, L., Narisetty, N. N., & Liang, F. (2019). Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, *114*, 1218–1231.

Gao, X., Pu, D. Q., Wu, Y., & Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of gaussian graphical model. *Statistica Sinica*, *22*, 1123–1146.

Ha, M. J., & Sun, W. (2014). Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation. *Biometrics*, *70*, 762–770.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*, 835–845.

Khondker, Z. S., Zhu, H., Chu, H., Lin, W., & Ibrahim, J. G. (2013). The bayesian covariance lasso. *Statistics and Its Interface*, *6*, 243.

Lian, H. (2011). Shrinkage tuning parameter selection in precision matrices estimation. *Journal of Statistical Planning and Inference*, *141*, 2839–2848.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, *34*, 1436–1462.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Patrascu, A., & Necoara, I. (2015). Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization. *Journal of Global Optimization*, *61*, 19–46.

Peng, J., Wang, P., Zhou, N., & Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, *104*, 735–746.

Raskutti, G., Yu, B., Wainwright, M. J., & Ravikumar, P. (2008). Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized mle. *Advances in Neural Information Processing Systems*, *21*.

Rossell, D., & Zwiernik, P. (2021). Dependence in elliptical partial correlation graphs. *Electronic Journal of Statistics*, *15*, 4236–4263.

Sustik, M. A., & Calderhead, B. (2012). *Glassofast: An efficient glasso implementation* (UTCS Technical Report No. TR-12-29 2012).

Vujačić, I., Abbruzzo, A., & Wit, E. (2015). A computationally fast alternative to cross-validation in penalized gaussian graphical models. *Journal of Statistical Computation and Simulation*, *85*, 3628–3640.

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, *7*, 867–886.

Wang, L., Ren, X., & Gu, Q. (2016). *Precision matrix estimation in high dimensional gaussian graphical models with faster rates*. In *Artificial intelligence and statistics* (pp. 177–185). PMLR.

Williams, D. R. (2020). *Beyond lasso: A survey of nonconvex regularization in gaussian graphical models*. https://doi.org/10.31234/osf.io/ad57p

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, *151*, 3–34.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, *94*, 19–35.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *38*, 894–942.

Zhou, S., van de Geer, S., & Bühlmann, P. (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*.

Zou, H. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Annals of Statistics*, *36*, 1509–1566.

How to cite this article: Carter, J. S., Rossell, D., & Smith, J. Q. (2024). Partial correlation graphical LASSO. *Scandinavian Journal of Statistics*, *51*(1), 32–63. https://doi.org/10.1111/sjos.12675

APPENDIX A. MEAN SQUARED ERROR OF LOGARITHMIC PENALTY

Section D outlines the derivation of the coordinate descent algorithm, and presents Algorithms SD.2–SD.2 to obtain the PCGLASSO solution for a sequence of penalization parameters and a given penalization parameter value, respectively. Section E provides supplementary results related to exchangeable inference. Section F provides empirical evidence that the difficulty of the inference problem depends only on Δ and not on θ.

Here we address the claim of Section 3 related to the mean squared error of logarithmic penalties in the $p = 1$ case. Specifically, we show that among penalty functions of the form $c\log(x)$ for constant $c \geq 0$ on the precision, choosing $c = 2$ minimizes the mean squared error of the estimate of the precision.

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

$$n\theta s \sim \chi^2_n,$$

and so

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

From this we get that

$$\mathbb{E}[s^{-1}] = \frac{n}{n-2}\theta,$$

$$\mathrm{Var}(s^{-1}) = \frac{2n^2}{(n-2)^2(n-4)}\theta^2.$$

Consider estimating θ via a penalized likelihood of the form

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

This can easily be shown to be maximized at

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

It follows that

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

$$\mathrm{Var}(\hat{\theta}) = \frac{2(1-\frac{2c}{n})^2 n^2}{(n-2)^2(n-4)}\theta^2,$$

and so

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= \mathrm{Var}(\hat{\theta}) + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2 \\ &= \theta^2\left(\frac{2(1-\frac{2c}{n})^2 n^2}{(n-2)^2(n-4)} + \left(\left(1 - \frac{2c}{n}\right)\left(\frac{n}{n-2}\right) - 1\right)^2\right).\end{aligned}$$

It can be shown that, for fixed θ and n, this function is minimized at $c = 2$, with the MSE at $c = 2$ being equal to $\frac{2\theta^2}{n-2}$.

APPENDIX B. PROOFS FOR SECTION 3

Proof of Proposition 1. Let S be some sample covariance matrix for which $\hat{\Theta}(S)$ is not diagonal and D be some diagonal matrix with nonzero 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 nonzero 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 maximize their corresponding penalized likelihoods, the derivatives of the penalized likelihood function (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

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

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

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

$$\text{pen}'_{ij}\left(\frac{\hat{\theta}_{ij}}{d_i d_j}\right) = d_i d_j \text{pen}'_{ij}(\hat{\theta}_{ij}). \tag{B1}$$

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

Second, if $\text{pen}'_{ij}(\hat{\theta}_{ij}) = k \neq 0$, then $\text{pen}'_{ij}\left(\frac{\hat{\theta}_{ij}}{d}\right) = dk$. Treating $\hat{\theta}_{ij}$, and therefore also k, as fixed, we denote by $x = \frac{\hat{\theta}_{ij}}{d}$. Then we have $\text{pen}'_{ij}(x) = \frac{\hat{\theta}_{ij}k}{x}$. It follows that $\text{pen}_{ij}(x) = \hat{\theta}_{ij}k\log(|x|) + c$ 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 $\hat{\theta}_{ii}$ maximizes the function:

$$\log(\theta_{ii}) - S_{ii}\theta_{ii} - \frac{2}{n}\text{pen}_{ii}(\theta_{ii}),$$

while $\tilde{\theta}_{ii}$ maximizes 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:

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

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 2 since the L_0 and logarithmic penalties are also symmetric PC-separable. ■

Proof of Proposition 2. 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 $\overline{\theta}^{1/2}\overline{\Delta}\overline{\theta}^{1/2}$ 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 $\overline{\Delta} = \text{sign}(D)\tilde{\Delta}\text{sign}(D)$ and $\overline{\theta} = D^2\tilde{\theta}$.

Since $\hat{\Theta}(S)$ maximizes the penalized likelihood at S, $\overline{\theta}, \overline{\Delta}$ must maximize

$$\log(\det(\Theta)) + \sum_i \left(\left(1 - \frac{2c}{n}\right)\log(\theta_{ii}) - s_{ii}\theta_{ii}\right) - \sum_{i \neq j}\left(s_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij} + \frac{2}{n}\text{pen}_{ij}(\Delta_{ij})\right), \tag{B2}$$

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

$$\log(\det(\Theta)) + \sum_i \left(\left(1 - \frac{2c}{n}\right)\log(\theta_{ii}) - d_i^2 s_{ii}\theta_{ii}\right) - \sum_{i \neq j}\left(d_i d_j s_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij} + \frac{2}{n}\text{pen}_{ij}(\Delta_{ij})\right). \tag{B3}$$

By substituting $\theta'_{ii} = d_i^2\theta_{ii}$ and $\Delta'_{ij} = \text{sign}(d_i d_j)\Delta_{ij}$ into (B3), and noting that pen_{ij} is symmetric about 0, we get

$$\log(\det(\Theta)) + \sum_i \left(\left(1 - \frac{2c}{n}\right)(\log(\theta'_{ii}) - \log(d_i^2)) - s_{ii}\theta'_{ii}\right) - \sum_{i \neq j}\left(s_{ij}\sqrt{\theta'_{ii}\theta'_{jj}}\Delta'_{ij} + \frac{2}{n}\text{pen}_{ij}(\Delta'_{ij})\right). \tag{B4}$$

Since $\log(d_i^2)$ is a constant, (B4) is of the same form as (B2) and they are maximized at the same point. Hence we have that $\overline{\Delta} = \text{sign}(D)\tilde{\Delta}\text{sign}(D)$ and $\overline{\theta} = D^2\tilde{\theta}$. ■

Proof of Proposition 3. Let π be a prior density as given in Proposition 3, S be some sample covariance and D some diagonal matrix with nonzero entries. Writing $L(\Theta|S)$ 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

$$\begin{aligned}
\pi(D\Theta D|S) &\propto L(D\Theta D|S)\pi(D\Theta D) \\
&\propto \det(\Delta)^{n/2}\prod_i (d_i^2\theta_{ii})^{\frac{n}{2}} \exp\left(-\frac{n}{2}\sum_{i,j} S_{ij}\sqrt{d_i^2\theta_{ii}d_j^2\theta_{jj}}\Delta_{ij}\right)
\end{aligned}$$

$$
\begin{aligned}
&\prod_i (d_i^2\theta_{ii})^{-c}\prod_{ij}\pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta\in\mathcal{S}_1)\\
&\quad= \det(\Delta)^{n/2}\prod_i (d_i^2\theta_{ii})^{\frac{n}{2}-c}\exp\left(-\frac{n}{2}\sum_{i,j} d_i d_j S_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}\right)\prod_{ij}\pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta\in\mathcal{S}_1),
\end{aligned}
\tag{B5}
$$

$$
\begin{aligned}
\pi(\Theta|DSD) &\propto L(\Theta|DSD)\pi(\Theta)\\
&\propto \det(\Delta)^{n/2}\prod_i \theta_{ii}^{\frac{n}{2}}\exp\left(-\frac{n}{2}\sum_{i,j} d_i d_j S_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}\right)\prod_i(\theta_{ii})^{-c}\prod_{ij}\pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta\in\mathcal{S}_1)\\
&= \det(\Delta)^{n/2}\prod_i \theta_{ii}^{\frac{n}{2}-c}\exp\left(-\frac{n}{2}\sum_{i,j} d_i d_j S_{ij}\sqrt{\theta_{ii}\theta_{jj}}\Delta_{ij}\right)\prod_{ij}\pi_{ij}(\Delta_{ij})\mathbb{I}(\Delta\in\mathcal{S}_1).
\end{aligned}
\tag{B6}
$$

For any measurable set A and $A_D = \{\Theta : D^{-1}\Theta D^{-1} \in A\}$ the probabilities in Definition 6 can be written as

$$
\begin{aligned}
\mathbb{P}_\pi(\Theta\in A|DSD) &= \int_A \pi(\Theta|DSD)\, d\Theta\\
&= \frac{\int_A L(\Theta|DSD)\pi(\Theta)\, d\Theta}{\int_S L(\Theta|DSD)\pi(\Theta)\, d\Theta},
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{P}_\pi(\Theta\in A_D|S) &= \int_{A_D} \pi(\Theta|S)\, d\Theta\\
&= \int_A \pi(D\Theta D|S)\, d\Theta\\
&= \frac{\int_A L(D\Theta D|S)\pi(D\Theta D)\, d\Theta}{\int_S L(D\Theta D|S)\pi(D\Theta D)\, d\Theta}.
\end{aligned}
$$

The result follows by noting that expression (B5) can be obtained by multiplying (B6) by the constant $\prod_i (d_i^2)^{\frac{n}{2}-c}$. ■

APPENDIX C. PROOFS FOR SECTION 6

Proof of Proposition 4. For a fixed θ, optimization of the penalized likelihood function (5) is equivalent to optimization 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 a sum of clearly concave functions. Hence the objective function is a sum of concave functions and is therefore concave. ■

APPENDIX D. COORDINATE DESCENT ALGORITHM

We present the coordinate descent algorithm we used to calculate PCGLASSO estimates in the simulated examples of this paper. Our aim is to find the values of Θ that maximize the objective function (5) for a sequence of penalty parameters $0 = \rho_0 < \rho_1 < \cdots < \rho_k$, that is, the regularization path. Algorithm 2, for which the coordinate descent algorithm D.2 is embedded, uses the previous estimate related to ρ_{i-1} as a starting point for the coordinate descent for ρ_i. This ensures that the coordinate descent is initialized at a point close to the maximum and aids convergence. For $\rho_0 = 0$ the algorithm is initialized at S^{-1}, or at $(S + \alpha I)^{-1}$ where I is the identity matrix if $n \leq p$. The matrix $S + \alpha I$ is guaranteed to be invertible and positive definite for any α. We also standardize the sample covariance S to have unit diagonals, before returning the estimates to the original scale. This has no effect on the estimated values due to the scale invariance of PCGLASSO, however it helps with the numerics of the coordinate descent.

Algorithm 3 is a standard blockwise coordinate descent algorithm which randomly cycles through the entries of Δ and maximizes 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 $q = \max\left\{\frac{2|\{\Delta^{(0)}_{ij} \neq 0 : i<j\}|}{p(p-1)}, \frac{2}{p(p-1)}\right\}$, the proportion of nonzero 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 3, results in Patrascu and Necoara (2015) and Wright (2015) suggest that convergence toward 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 3. However, we prefer cycling through the indices without replacement since this provides a more simple and clear stopping rule for the algorithm. Algorithm 3 assesses the convergence after updating each entry of Δ exactly once, so

Algorithm 2. PCGLASSO regularization path

```
Input  : Sample covariance $S$, sequence of regularization parameters
         $0 = \rho_0 < \rho_1 < \cdots < \rho_k$ and optimization convergence threshold $\epsilon$.
Output: Sequence of estimates $\Theta_0, \dots, \Theta_k$ corresponding to $\rho_0, \dots, \rho_k$.
 1. standardize the sample covariance $\tilde{S} = \mathrm{diag}(S)^{-1/2} S \mathrm{diag}(S)^{-1/2}$.
 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 $i = 1, \dots, k$, run Algorithm 3 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 $\Theta_i = \mathrm{diag}(S)^{-1/2} \tilde{\Theta}_i \mathrm{diag}(S)^{-1/2}$ for $i = 0, 1, \dots, k$.
```

Algorithm 3. Blockwise coordinate descent

```
Input  : Sample covariance $S$ with unit diagonal, penalty parameter $\rho$, start point $\Theta^{(0)}$
         and optimization convergence threshold $\epsilon$.
Output: A matrix $\Theta$ providing a local maximum of (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 $\Delta_{ij}, \theta_{ii}, \theta_{jj}$ maximize

        $f(\Delta, \theta) = \log(\det(\Delta)) + \left(1 - \frac{4}{n}\right) \sum_i \log(\theta_{ii}) - \mathrm{tr}\left(S\theta^{1/2}\Delta\theta^{1/2}\right) - \rho\|\Delta\|_1,$

    subject to

        $\Delta \in \mathcal{S}_1,$
        $\Delta_{k_1k_2} = \Delta^{(1)}_{k_1k_2},$ 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 nonzero 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.
```

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.

As a final note about Algorithm 3, Step 2 maximizing (5) with respect to $\Delta_{ij}, \theta_{ii}, \theta_{jj}$ while all other variables are held fixed is nontrivial due to the nonsmoothness of the objective function. The remainder of this section will focus on solving this maximization problem. 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

$$c_n = 1 - \frac{4}{n},$$

$$c_{12} = S_{ij},$$

$$c_1 = \sum_{k \neq i,j} S_{ik}\Delta_{ik}\sqrt{\theta_{kk}},$$

$$c_2 = \sum_{k \neq i,j} S_{jk}\Delta_{jk}\sqrt{\theta_{kk}}.$$

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 maximization problem can then be expressed as

$$\begin{aligned} \max_{x,y_1,y_2} \quad & f(x,y_1,y_2) \\ \text{s.t.} \quad & x \in (l,u) \\ & y_1,y_2>0. \end{aligned} \tag{D1}$$

We denote the partial derivatives of f by

$$f_x(x,y_1,y_2) = \frac{2ax+b}{ax^2+bx+c} - 2c_{12}y_1y_2 - 2\rho\,\text{sign}(x), \quad x \neq 0,$$

$$f_{y_1}(x,y_1,y_2) = 2c_ny_1^{-1} - 2y_1 - 2c_{12}xy_2 - 2c_1,$$

$$f_{y_2}(x,y_1,y_2) = 2c_ny_2^{-1} - 2y_2 - 2c_{12}xy_1 - 2c_2,$$

To solve this problem we consider separately the cases $c>0$ and $c\leq 0$.

D.1 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

$$x = \frac{c_ny_1^{-1}-y_1-c_1}{c_{12}y_2}, \tag{D2}$$

and setting $f_{y_2}(x, y_1, y_2) = 0$ along with (D2) gives

$$y_2 = \frac{1}{2}\left(-c_2 \pm \sqrt{c_2^2 + 4(y_1^2 + c_1 y_1)}\right). \tag{D3}$$

Using (D2) and (D3) 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 (D2) and (D3).

The constraint $x < u$ results in some condition on the following quartic which we refer to as $q(y_1)$

$$\begin{aligned}&\left(1 - \frac{1}{u^2 c_{12}^2}\right) y_1^4 + \left(c_1 - \frac{2c_1}{u^2 c_{12}^2} + \frac{c_2}{u c_{12}}\right) y_1^3 \\ &+ \left(\frac{2c_n}{u^2 c_{12}^2} - \frac{c_1^2}{u^2 c_{12}^2} + \frac{c_1 c_2}{u c_{12}}\right) y_1^2 \\ &+ \left(\frac{2c_1 c_n}{u^2 c_{12}^2} - \frac{c_2 c_n}{u c_{12}}\right) y_1 - \frac{c_n^2}{u^2 c_{12}^2}.\end{aligned} \tag{D4}$$

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 (D3) 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 (D3) 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 (D2) implies that

$$\operatorname{sign}(y_1^2 + c_1 y_1 - c_n) = -\operatorname{sign}(c_{12}).$$

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 optimization problem is simpler and so the details of this case are omitted.

For y_2 to take a real value in (D3) we must have $4y_1^2 + 4c_1y_1 + c_2^2 \geq 0$ 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 (D3), 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 (D3). For all other situations one must only consider the positive root.

Now combining the constraint $x < u$ with (D2) and (D3), one obtains the inequality

$$\frac{2}{uc_{12}}\left(c_n y_1^{-1} - y_1 - c_1\right) + c_2 < \sqrt{c_2^2 + 4(y_1^2 + c_1y_1)},$$

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, (D3) and (D2) 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.

D.2 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 $b > 0$, (D2) 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.

APPENDIX E. SUPPLEMENTARY RESULTS FOR SECTION 4

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) $S_{k_1k_2}/\hat{\theta}_{k_2k_2}^{-1/2} = S_{k_1k_3}/\hat{\theta}_{k_3k_3}^{-1/2}$.
(C2) $\hat{\Delta}_{k_2j} = \hat{\Delta}_{k_3j}$ for all $j \notin \{k_1, k_2, k_3\}$.

Proposition 5. *Under conditions (C1) and (C2) 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 maximizes the function

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

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) and (C2).

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

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

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

$$2\hat{\theta}_{11}^{1/2}s_{12}\hat{\theta}_{22}^{1/2}(\Delta_{12} + \Delta_{13}) + c, \tag{E1}$$

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

$$(\Delta_{12}, \Delta_{13}, \hat{\Delta}_{14}, \ldots, \hat{\Delta}_{1p})\hat{\Delta}_{2:p,2:p}^{-1}(\Delta_{12}, \Delta_{13}, \hat{\Delta}_{14}, \ldots, \hat{\Delta}_{1p})^T, \tag{E2}$$

is symmetric in $(\Delta_{12}, \Delta_{13})$. To ease notation let $A = \hat{\Delta}_{2:p,2:p}^{-1}$. Note that under Condition (C2),

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

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, (E2) is equal to

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

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. ■

Note that because the log-likelihood is a convex function, and therefore has a unique maximum, symmetry in $(\Delta_{k_1k_2}, \Delta_{k_1k_3})$ implies that the MLE will estimate these two partial correlations to be equal.

Corollary 1. *Under conditions (C1) and (C2) any penalized 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 5, noting that $\mathrm{Pen}(\theta, \Delta) = \sum_i \mathrm{pen}_{ii}(\theta_{ii}) + \sum_{i \neq j} \mathrm{pen}(\Delta_{ij})$ is symmetric in $(\Delta_{12}, \Delta_{13})$. ■

Corollary 2. *Under conditions (C1) and (C2) a penalized 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 5 the penalized likelihood is symmetric if and only if $\mathrm{pen}_{k_1k_2}\left(\sqrt{\hat{\theta}_{k_1k_1}\hat{\theta}_{k_2k_2}}\Delta_{k_1k_2}\right) + \mathrm{pen}_{k_1k_3}\left(\sqrt{\hat{\theta}_{k_1k_1}\hat{\theta}_{k_3k_3}}\Delta_{k_1k_3}\right)$ 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. ■

APPENDIX F. INFERENCE DIFFICULTY

Consider the $p = 2$ dimensional case with partial correlation matrix and diagonal entries

$$\Delta = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \quad \theta = \begin{pmatrix} x \\ x \end{pmatrix}.$$

Here the partial correlation is fixed but we consider varying (but equal) diagonal entries.

Suppose we use the MLE S^{-1} as an estimator for the precision matrix (and therefore $\frac{S_{12}^{-1}}{\sqrt{S_{11}^{-1}S_{22}^{-1}}}$ as an estimator for the partial correlation $\Delta_{12} = -0.5$). In Figure F1 we show the mean squared error (MSE) for the partial correlation estimator when the sample size is $n = 10$ for various values of x (here the mean squared error has been approximated by sampling 100,000 times). We see that

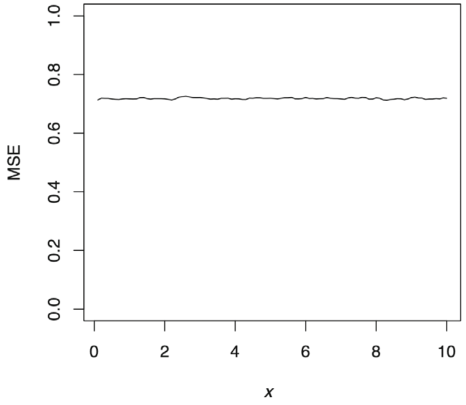

FIGURE F1 Mean squared error of the maximum likelihood estimator (MLE) estimator of the partial correlation when $p = 2$, $n = 10$, the partial correlation is fixed $\Delta_{12} = -0.5$ and the diagonal entries are equal to $\theta_{11} = \theta_{22} = x$.

the MSE does not depend on x—hence the difficulty of estimating Δ_{12} or detecting nonzero Δ_{12} does not depend on the diagonal entries.

Now instead consider the precision matrix

$$\Theta = \begin{pmatrix} x & -0.5 \\ -0.5 & x \end{pmatrix}.$$

Now the off-diagonals are fixed and again the diagonal entries are allowed to vary (with $x > 0.5$).

In this case, the MSE of the estimator for the partial correlation increases with x (see left panel of Figure F2). Furthermore, the MSE of the estimator S^{-1}_{12} of θ_{12} also increases with x, even though θ_{12} remains constant (see right panel of Figure F2). This shows that the inference task of estimating θ_{ij} and detecting nonzero θ_{ij} does depend on the diagonal entries.

This shows that, in the $p = 2$ case, the difficulty of detecting a nonzero partial correlation (or equivalently a nonzero off-diagonal) and estimating its value is better expressed in the θ, Δ parameterization. In particular, the difficulty only depends on Δ and not on θ. On the other hand, with the Θ parameterization, the difficulty depends on the whole matrix.

Now consider the $p = 10$ case with a star graph structure—that is $\Delta_{1i} = \Delta_{i1} = -x$ for all i, while all other $\Delta_{ij} = 0$. We also fix $\theta_{ii} = 1$ for all i. For this setting we find the mean MCC of GLASSO (with penalty parameter chosen to minimize the BIC) over 100 repetitions when the sample size is $n = 50$ and for varying x (note that when GLASSO returns a diagonal matrix we assign this an MCC of 0). The results of this can be found in the left panel of Figure F3. We see that the model selection deteriorates as the non-zero partial correlations get smaller in magnitude.

Now consider the same setting but with fixed partial correlations $\Delta_{1i} = \Delta_{i1} = -1/\sqrt{10}$ for all i and varying diagonals $\theta_{ii} = x$ for all i. In the center panel of Figure F3 we see that the model selection of GLASSO remains constant for different x. This will also trivially hold for PCGLASSO due to the scale invariance property of Proposition 2.

Finally consider the Θ parameterization where the off-diagonals are fixed $\theta_{1i} = \theta_{i1} = -1/\sqrt{10}$ for all i and varying diagonals $\theta_{ii} = x$ for all i. In this case the model selection deteriorates as x increases (see the right panel of Figure F3).

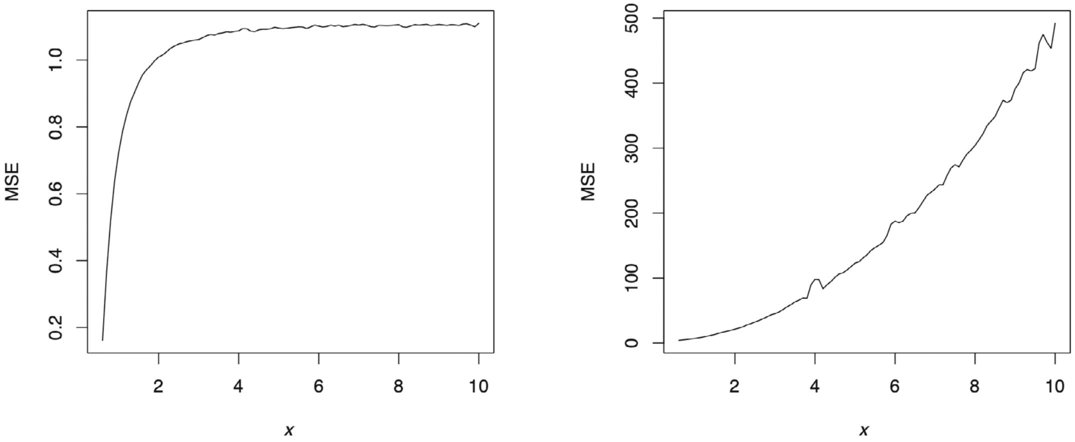

FIGURE F2 Mean squared error of the maximum likelihood estimator (MLE) estimator of the partial correlation (left) and off-diagonal (right) when $p = 2$, $n = 10$, the off-diagonal is fixed $\theta_{12} = -0.5$ and the diagonal entries are equal to $\theta_{11} = \theta_{22} = x$.

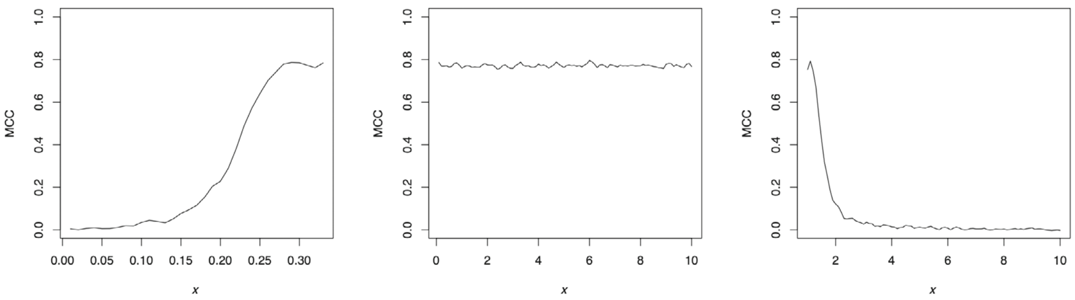

FIGURE F3 Matthews correlation coefficient in the star example for varying partial correlation Δ_{ij} (left), varying diagonal with fixed Δ_{ij} (centre) and varying diagonal with fixed θ_{ij} (right).

These three examples demonstrate that the model selection depends on the magnitude of the Δ_{ij} and that, for fixed partial correlations, the model selection does not depend on the diagonals θ_{ii}. It also shows that the magnitude of the Δ_{ij} is a better measure of the problem difficulty than the magnitude of the θ_{ij}.