The Log-Shift Penalty for Adaptive Estimation of Multiple Gaussian Graphical Models

Yuancheng Zhu
University of Chicago

Rina Foygel Barber
University of Chicago

Abstract

Sparse Gaussian graphical models characterize sparse dependence relationships between random variables in a network. To estimate multiple related Gaussian graphical models on the same set of variables, we formulate a hierarchical model, which leads to an optimization problem with a nonconvex log-shift penalty function. We show that under mild conditions the optimization problem is convex despite the inclusion of a nonconvex penalty, and derive an efficient optimization algorithm. Experiments on both synthetic and real data show that the proposed method is able to achieve good selection and estimation performance simultaneously, because the nonconvexity of the log-shift penalty allows for weak signals to be thresholded to zero without excessive shrinkage on the strong signals.

1 INTRODUCTION AND BACKGROUND

For a set of variables $X_1, \dots, X_p$, a graphical model is commonly used to reflect sparse dependence structure among the variables. The presence of an edge (i, j) reflects that variables X_i and X_j are dependent even after controlling for the effects of the remaining variables. If $X = (X_1, \dots, X_p) \sim N(\mu, \Sigma)$, the resulting model is known as a Gaussian graphical model (GGM), and in this case the edges (i.e. conditional dependencies) correspond to nonzero entries in the precision matrix, $\Omega = \Sigma^{-1}$. The log-likelihood for Ω after observing n i.i.d. draws of X is given by

$$L(\Omega) = \frac{n}{2} \log\det(\Omega) - \frac{n}{2}\langle S, \Omega\rangle \ ,$$

where S is the sample covariance of the n i.i.d. observations. These types of models arise in a wide range of applications, including genetics (modeling interactions among gene expression levels), finance (finding interactions between different stock prices), and social networks (modeling relationships among people, spread of information or disease, etc).

In a high-dimensional setting, where we observe n i.i.d. realizations of X for a sample size $n < p$, the sparsity of the precision matrix allows us to accurately estimate the distribution of X even though $\Sigma = \mathsf{Cov}(X)$ is in general not identifiable from $n < p$ samples. A well-studied convex approach to finding $\Omega = \Sigma^{-1}$ is the graphical Lasso [5], which calculates[1]

$$\widehat{\Omega}_{\mathsf{glasso}} = \arg\min_{\Omega \succeq 0} \Big\{ - L(\Omega) + \gamma \sum_{i<j} |\Omega_{ij}| \Big\} \ . \qquad (1)$$

The penalty term promotes sparsity—due to the shrinkage on the off-diagonal entries of Ω, many of the Ω_{ij}'s (for $i \neq j$) will be zero when γ is sufficiently large. Under some conditions, the graphical Lasso is consistent for edge selection in sparse models, even at sample size $n \ll p$ [13].

Multiple graphs In some applications, we may have multiple sets of observations with related (but not necessarily identical) covariance structures, for instance when the same variables are measured across different settings (such as gene expression levels in healthy vs in cancerous tissues [3] or across different phases of an organism's life cycle [9]). Suppose that we observe data from K different GGMs with similar sparsity structures, and would like to estimate the K precision matrices $\Omega^{(1)}, \dots, \Omega^{(K)}$ jointly. Let $L_k(\Omega^{(k)})$ be the log-likelihood for the kth data set given by

$$L_k(\Omega^{(k)}) = \frac{n_k}{2} \log\det(\Omega^{(k)}) - \frac{n_k}{2}\langle S^{(k)}, \Omega^{(k)}\rangle$$

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38. Copyright 2015 by the authors.

[1] In some works in the literature, $\|\Omega\|_1 = \sum_{ij} |\Omega_{ij}|$ is penalized, i.e. the diagonal elements are not excluded from the penalty, but we exclude them to facilitate comparison with our work.

for kth sample size n_k and kth sample covariance matrix $S^{(k)}$. Danaher et al. [3] propose the group graphical Lasso:

$$\widehat{\mathbf{\Omega}}_{\mathsf{GGL}} = \arg\min_{\mathbf{\Omega}\in\mathcal{S}_p} \Big\{ -\sum_k L_k(\Omega^{(k)}) + \gamma \sum_{i<j} \big[\nu\|\mathbf{\Omega}_{ij}\|_1 + (1-\nu)\|\mathbf{\Omega}_{ij}\|_2\big]\Big\}, \tag{2}$$

where $\mathcal{S}_p$ is the feasible set of positive semidefinite matrix sequences,

$$\mathcal{S}_p = \Big\{\mathbf{\Omega} = (\Omega^{(1)},\dots,\Omega^{(K)}) \;:\; \Omega^{(k)} \in \mathbb{R}^{p\times p}, \Omega^{(k)} \succeq 0\Big\},$$

and $\mathbf{\Omega}_{ij} = (\Omega^{(1)}_{ij},\dots,\Omega^{(K)}_{ij})$ is the vector of coefficients at position (i,j) across the K settings. If $\nu = 1$, the solution $\widehat{\mathbf{\Omega}}_{\mathsf{GGL}}$ reduces to performing a graphical Lasso on each data set $k = 1,\dots,K$; no information is shared across the K tasks. At the other extreme, for $\nu = 0$, the penalty $\sum_{i<j}\|\mathbf{\Omega}_{ij}\|_2$ ensures identical sparsity patterns across the K estimated precision matrices.

A nonconvex approach For the single graph setting ($K = 1$), recent work by Wong et al. [15] proposes an adaptive, nonconvex approach, defined by a hierarchical model for each Ω_{ij}:

$$\begin{aligned} \tau_{ij} &\propto 1/\tau_{ij} \text{ for all } i<j \text{ (an improper prior)}, \qquad (3)\\ \Omega_{ij}|\tau_{ij} &\sim \mathcal{N}(0,\tau_{ij}) \text{ for all } i<j,\\ X|\mu,\Omega &\sim \mathcal{N}\big(\mu,\Omega^{-1}\big) \text{ i.i.d. for each observation.}\end{aligned}$$

This hierarchical model can be viewed as a graphical model version of the Bayesian Lasso introduced by Park and Casella [11]. Marginalizing over τ_{ij}, this induces a marginal (improper) density $\propto 1/|\Omega_{ij}|^2$, leading to the MAP estimation problem

$$\widehat{\Omega}_{\mathsf{adaptive}} = \arg\min_{\Omega\succeq 0}\Big\{-L(\Omega) + \sum_{i<j}\log(|\Omega_{ij}|^2)\Big\}. \tag{4}$$

This procedure is adaptive because, due to the concavity of the log penalty, large entries Ω_{ij} suffer less shrinkage when estimated, as compared to an ℓ_1-norm penalty like the graphical Lasso. Empirically, Wong et al. [15] find that the adaptive nonconvex penalty leads to improvements relative to the graphical Lasso (1) in terms of accurate estimation and support recovery. However, the optimization problem (4) is in general nonconvex and may have local minima.

Contributions In the work presented here, we formulate a hierarchical model for multiple graphs, and derive an optimization problem corresponding to finding the maximum a posteriori (MAP) estimate for $\mathbf{\Omega}$. The resulting optimization problem combines a likelihood term with a nonconvex penalty, leading to reduced shrinkage on edges with strong signals (thus improving over convex-penalty methods). Crucially, even with the nonconvex penalty, our optimization problem is convex under some mild conditions, thus avoiding issues with local minima. Furthermore, we find that the optimization speedup results of Danaher et al. [3] extend to our method. Empirically, our method is able to simultaneously identify the nonzero edges in a graph (model selection) and estimate the parameters on these edges—this is a strong advantage of our nonconvex penalty, which is able to produce a sparse solution while not imposing strong shrinkage on large nonzero estimated values, while convex-penalty methods generally cannot achieve both at the same tuning parameter value.

Outline The remainder of this paper is organized as follows. We introduce our method in Section 2, which gives a hierarchical model for the K linked GGMs, and derives an objective function to find the maximum a posteriori (MAP) estimate for $\mathbf{\Omega} = (\Omega^{(1)},\dots,\Omega^{(K)})$. We discuss the sparsity and shrinkage properties of our method, in particular as compared to the group graphical Lasso, in Section 2.2. In Section 3 we discuss optimization for the objective function defined by our method, and in particular find conditions that lead to a convex optimization problem that can be split into smaller subproblems (connected components of the graphs); proofs for the results in this section can be found in the Supplementary Materials. We present experiments on simulated data and on two real data sets (stock price data and bikeshare data) in Section 4. Finally, we conclude with a brief discussion of our work and of future directions in Section 5.

2 METHODOLOGY

2.1 A Hierarchical Model for Multiple GGMs

Consider the following hierarchical models for K Gaussian graphical models with p nodes each:

$$\begin{aligned} \tau_{ij} &\sim \text{InverseGamma}(\alpha,\beta) \text{ for all } i<j, \qquad (5)\\ \Omega^{(k)}_{ij}|\tau_{ij} &\sim \text{Laplace}\,(\tau_{ij}) \text{ for all } k, \text{ for all } i<j,\\ X^{(k)}|\mu^{(k)},\Omega^{(k)} &\sim \mathcal{N}\big(\mu^{(k)},(\Omega^{(k)})^{-1}\big) \text{ for all } k.\end{aligned}$$

We place a flat (improper) prior on $\mu^{(k)}$ and on the diagonal entries $\Omega^{(k)}_{ii}$. Of course, we must require $\Omega^{(k)} \succeq 0$ for each k. We may also choose to allow improper priors for τ_{ij} by allowing α and/or β to be zero.

This hierarchical model characterizes our prior belief regarding shared structure across the K graphs. The common structure across the graphs is governed by the shared parameter τ_{ij} for the weights on the same edge in different graphs. The hyperparameters α and β control the magnitude and the variation of the τ_{ij}'s and thus the sparsity pattern of the graphs.

Marginal distribution of $\boldsymbol{\Omega}$ given τ We now calculate the marginal prior density of $\boldsymbol{\Omega}$:

$$\begin{aligned}
p(\boldsymbol{\Omega}) &\propto \mathbb{1}_{\boldsymbol{\Omega}\in\mathcal{S}_p}\cdot\prod_{i<j}\int_{\tau_{ij}}\left[\prod_{k=1}^{K}p(\Omega_{ij}^{(k)}|\tau_{ij})\right]p(\tau_{ij})\ \mathsf{d}\tau_{ij}\\
&\propto \mathbb{1}_{\boldsymbol{\Omega}\in\mathcal{S}_p}\cdot\prod_{i<j}\int_{\tau_{ij}}\left[\prod_{k=1}^{K}\tau_{ij}^{-1}e^{-\frac{|\Omega_{ij}^{(k)}|}{\tau_{ij}}}\right]\tau_{ij}^{-\alpha-1}e^{-\frac{\beta}{\tau_{ij}}}\ \mathsf{d}\tau_{ij}\\
&= \mathbb{1}_{\boldsymbol{\Omega}\in\mathcal{S}_p}\cdot\prod_{i<j}\int_{\tau_{ij}}\tau_{ij}^{-K-\alpha-1}e^{-(\beta+\|\boldsymbol{\Omega}_{ij}\|_1)/\tau_{ij}}\ \mathsf{d}\tau_{ij}\\
&\propto \mathbb{1}_{\boldsymbol{\Omega}\in\mathcal{S}_p}\cdot\prod_{i<j}(1+\|\boldsymbol{\Omega}_{ij}\|_1/\beta)^{-(\alpha+K)}\ , \qquad (6)
\end{aligned}$$

where the last step is obtained by marginalizing over τ_{ij} and dividing by the constant $\beta^{-(\alpha+K)}$. When $K = 1$, even though our hierarchical model takes a different form than the model (3) proposed by [15], we obtain the same marginal distribution of $\boldsymbol{\Omega}$ when we take $\alpha = 1$ and $\beta \to 0$. However, we will show later on that choosing nonzero β will allow for a convex optimization problem.

The posterior MAP Combining the marginal prior on $\boldsymbol{\Omega}$ (6) with the log-likelihoods $L_k(\Omega^{(k)})$ from the K data sets, we would like to calculate the maximum a posteriori (MAP) estimate:

$$\widehat{\boldsymbol{\Omega}} = \arg\min_{\boldsymbol{\Omega}\in\mathcal{S}_p}\Big\{-\sum_k L_k(\Omega^{(k)})+\gamma\sum_{i<j}\beta\log\big(1+\frac{\|\boldsymbol{\Omega}_{ij}\|_1}{\beta}\big)\Big\}, \qquad (7)$$

where $\gamma = \frac{\alpha+K}{\beta}$ (we introduce this reparametrization for later convenience). This penalized likelihood function combines a convex negative-log-likelihood term with a nonconvex "log-shift" penalty. While the underlying hierarchical model requires $\gamma\beta \geq K$ by construction, we relax this to $\gamma \geq 0$.

A generalization We can also consider replacing $\|\boldsymbol{\Omega}_{ij}\|_1$ in (7) with any convex regularizer $f(\boldsymbol{\Omega}_{ij})$, which leads to the optimization problem

$$\widehat{\boldsymbol{\Omega}} = \arg\min_{\boldsymbol{\Omega}\in\mathcal{S}_p} F(\boldsymbol{\Omega}) \qquad (8)$$

where

$$F(\boldsymbol{\Omega}) := -\sum_k L_k(\Omega^{(k)}) + \gamma\sum_{i<j}\beta\log(1+f(\boldsymbol{\Omega}_{ij})/\beta)\ .$$

As an important example, we can consider a (sparse) group Lasso penalty on each $\boldsymbol{\Omega}_{ij}$:

$$f(\boldsymbol{\Omega}_{ij}) = \nu\|\boldsymbol{\Omega}_{ij}\|_1 + (1-\nu)\|\boldsymbol{\Omega}_{ij}\|_2\ .$$

In fact, the penalized likelihood optimization problem in (8) can be motivated by a generalization of our hierarchical model in (5). Consider

$$\begin{aligned}
\tau_{ij} &\sim \text{InverseGamma}(\alpha,\beta) \text{ for all } i<j, \qquad (9)\\
\boldsymbol{\Omega}_{ij}|\tau_{ij} &\propto \tau_{ij}^{-K}e^{-f(\boldsymbol{\Omega}_{ij})/\tau_{ij}} \text{ for all } i<j,\\
X^{(k)}|\Omega^{(k)} &\sim \mathcal{N}\big(\mu^{(k)},(\Omega^{(k)})^{-1}\big) \text{ for all } k.
\end{aligned}$$

As before, we place a flat prior on $\mu^{(k)}$ and on the diagonal entries $\Omega_{ii}^{(k)}$, and require $\Omega^{(k)} \succeq 0$. Next, marginalizing over τ

$$\begin{aligned}
p(\boldsymbol{\Omega}) &\propto \mathbb{1}_{\boldsymbol{\Omega}\in\mathcal{S}_p}\cdot\prod_{i<j}\int_{\tau_{ij}}p(\boldsymbol{\Omega}_{ij}|\tau_{ij})p(\tau_{ij})\ \mathsf{d}\tau_{ij}\\
&\propto \mathbb{1}_{\boldsymbol{\Omega}\in\mathcal{S}_p}\cdot\prod_{i<j}\int_{\tau_{ij}}\tau_{ij}^{-K}e^{-f(\boldsymbol{\Omega}_{ij})/\tau_{ij}}\cdot\tau_{ij}^{-\alpha-1}e^{-\beta/\tau_{ij}}\ \mathsf{d}\tau_{ij}\\
&\propto \mathbb{1}_{\boldsymbol{\Omega}\in\mathcal{S}_p}\cdot\prod_{i<j}(1+f(\boldsymbol{\Omega}_{ij})/\beta)^{-(\alpha+K)}\ .
\end{aligned}$$

Combining this with the likelihood terms, and setting $\gamma = \frac{\alpha+K}{\beta}$ as before, yields the optimization problem (8).

2.2 Sparsity and Shrinkage of $\boldsymbol{\Omega}$

We next examine the effects of the parameters γ and β in the log-shift objective function (8), which arise from α and β in the hierarchical model (9). To understand their role in inducing sparsity and shrinkage in $\boldsymbol{\Omega}$, we first consider the function

$$g_\beta(x) = \beta\log(1+|x|/\beta)\ .$$

In a sparse regression setting, this type of penalty function has been studied by Candès et al. [1] and others in the context of reweighted ℓ_1 minimization, and was found to preserve the desirable sparsity properties of ℓ_1 regularization while reducing the amount of shrinkage on large coefficients.

The penalty function $g_\beta(x)$ behaves like a ℓ_1 penalty when $|x|/\beta \approx 0$, which we can see by taking a local linear approximation to the log function:

$$\log(1+|x|/\beta) \approx |x|/\beta\ \Rightarrow\ g_\beta(x) \approx |x|\ .$$

On the other hand, as $|x|/\beta$ grows large, the concavity of the log function becomes apparent, and therefore there is less shrinkage on large values of x (see Figure 1).

Next, we return to our prior distribution on $\boldsymbol{\Omega}$. Comparing the penalized likelihood function (8) with our calculations with $g_\beta(\cdot)$ above, we can interpret the parameters β and γ in (8) as follows:

- γ controls the amount of penalization on $\boldsymbol{\Omega}$, and thus the sparsity level of the solution.
- β controls the nonconvexity of the penalty, with small β yielding reduced shrinkage in the estimate of $\boldsymbol{\Omega}$ (but possible nonconvexity of the objective function), while $\beta \to \infty$ causes the penalty term to approach $\gamma\sum_{i<j}f(\boldsymbol{\Omega}_{ij})$.

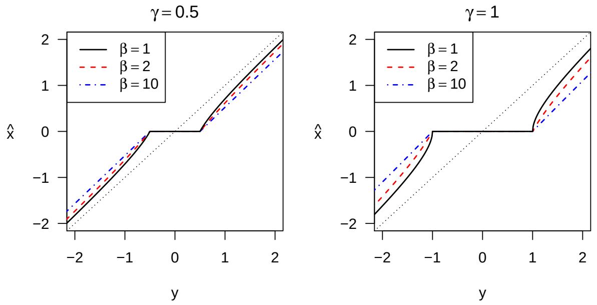

Figure 1: Sparsity and shrinkage behavior of the solution $\hat{x} = \arg\min\left\{\frac{1}{2}(y-x)^2 + \gamma g_\beta(x)\right\}$. Note that γ affects the point at which the solution is thresholded to zero, while β controls the nonconvexity (and therefore, the shrinkage) for nonzero solutions.

2.3 Other Related Work

The graphical Lasso [5] and group graphical Lasso [3] methods, discussed above in (1) and (2), both propose estimation of $\mathbf{\Omega}$ via a convex penalty. Our method may be viewed as a generalization of the group graphical Lasso, which is obtained by setting $\beta = \infty$ in our log-shift penalty.

Turning to nonconvex methods, in addition to Wong et al. [15]'s model described in (3) above, we are aware of several other methods using nonconvex regularization, all of which allow for reduced shrinkage on large entries, but may potentially lead to nonconvex optimization problems. First, for estimation of a single graph, Fan et al. [4] apply an adaptive Lasso (reweighted ℓ_1) penalty to the GGM setting, optimizing

$$\arg\min_{\Omega\succeq 0}\left\{-L(\Omega) + \gamma\sum_{i<j}|\Omega_{ij}|/|\widetilde{\Omega}_{ij}|^\alpha\right\},$$

where $\widetilde{\Omega}$ is some initial estimate of Ω. In fact, for $\alpha = 1$, this reweighted ℓ_1 penalty is can be viewed as a single iteration towards solving Wong et al. [15]'s MAP estimation problem (4), although this is not the approach taken in [15] (see [1] for the sparse regression setting). Fan et al. [4] also examine a SCAD penalty on each $|\Omega_{ij}|$, which behaves similarly.

In the multiple graph setting, Guo et al. [7] propose the optimization problem

$$\widehat{\mathbf{\Omega}}_{\mathsf{sqrt}} = \arg\min_{\mathbf{\Omega}\in\mathcal{S}_p}\left\{-\sum_k L_k(\Omega^{(k)}) + \gamma\sum_{i<j}\sqrt{\|\mathbf{\Omega}_{ij}\|_1}\right\}. \tag{10}$$

This nonconvex penalty encourages similar sparsity patterns in the K graphs, but does not allow for tuning the amount of nonconvexity or the balance between shared support vs different support.

Finally, in a recent work, Zhu et al. [17] propose a nonconvex method for simultaneous estimation of multiple GGMs by penalizing the log likelihood with the following penalty function

$$\gamma_1\sum_k\sum_{i<j}J_\tau(|\Omega^{(k)}_{ij}|) + \gamma_2\sum_{k\neq k'}\sum_{i<j}J_\tau(|\Omega^{(k)}_{ij} - \Omega^{(k')}_{ij}|),$$

where $J_\tau(z) = \min(|z|,\tau)$ is the truncated ℓ_1-norm penalty. This optimization problem allows for tuning with the three parameters, but regardless of tuning parameter values, the objective function is nonconvex due to the shape of the truncated ℓ_1-norm penalty.

3 CONVEXITY AND OPTIMIZATION

In this section, we derive a simple condition on the parameters β and γ in our log-shift method (8) that guarantees the convexity of the objective function $F(\mathbf{\Omega})$ over a bounded set. We then develop a majorization-minimization algorithm for finding the global minimum, and a preprocessing step where the graphs are split into connected components, allowing for smaller optimization problems that can be solved in parallel. While we are primarily interested in regularizers of the form $f(\cdot) = \nu\|\cdot\|_1 + (1-\nu)\|\cdot\|_2$, our results apply more generally to any convex regularizer $f(\cdot)$.

3.1 Convexity

To ensure that $F(\mathbf{\Omega})$ is convex, we will place a bound on $\mathbf{\Omega}$ to obtain strong convexity of the likelihood term, while placing a lower bound on β to control the nonconvexity of the penalty term. The result below may be viewed as an application of Loh and Wainwright [10]'s results on nonconvex regularizers.

The condition that we require on $\mathbf{\Omega}$ is mild. For any $b = (b_1,\dots,b_K)\in\mathbb{R}^K_+$, define

$$\mathcal{S}_p(b) = \left\{\mathbf{\Omega}\in\mathcal{S}_p\ :\ \|\Omega^{(k)}\|_{\mathsf{op}} \le b_k \text{ for } k = 1,\dots,K\right\},$$

where $\|\cdot\|_{\mathsf{op}}$ is the matrix operator norm (i.e. the largest singular value). This is a reasonable nondegeneracy condition on the K graphical models underlying the data.

Theorem 1. *If $f(\cdot)$ is convex, nonnegative, and L-Lipschitz, and if*

$$\beta \ge \frac{\gamma L^2}{2}\cdot\max_k\frac{b_k^2}{n_k}, \tag{11}$$

then $F(\mathbf{\Omega})$ is convex over $\mathbf{\Omega}\in\mathcal{S}_p(b)$. If (11) is satisfied with a strict inequality, then we obtain strict convexity.

We note that when the sample sizes n_k are all large, the condition (11) allows β to be very small, that is, allows the penalty to be highly nonconvex, as desired to avoid excessive shrinkage on strong signals. The proof of this theorem, given in the Supplementary Materials, simply shows that the strong convexity of the likelihood term in $F(\mathbf{\Omega})$ is sufficient to counterbalance the concavity of the log penalty.

3.2 Optimization via Majorization-Minimization

To minimize $F(\mathbf{\Omega})$ we use majorization-minimization [8]. Let $\widetilde{\mathbf{\Omega}}$ be our current estimate of $\widehat{\mathbf{\Omega}}$. Since $\log(\cdot)$ is concave, we bound $\log(1 + f(\mathbf{\Omega}_{ij})/\beta)$ by the linear approximation centered at $\widetilde{\mathbf{\Omega}}_{ij}$:

$$\log\big(1 + f(\mathbf{\Omega}_{ij})/\beta\big) \\ \leq \log\big(1 + f(\widetilde{\mathbf{\Omega}}_{ij})/\beta\big) + \frac{f(\mathbf{\Omega}_{ij})/\beta - f(\widetilde{\mathbf{\Omega}}_{ij})/\beta}{1 + f(\widetilde{\mathbf{\Omega}}_{ij})/\beta} \ .$$

Then the objective function is bounded as

$$F(\mathbf{\Omega}) \leq \underbrace{-\sum_k L_k(\Omega^{(k)}) + \gamma \sum_{i<j} \frac{f(\mathbf{\Omega}_{ij})}{1 + f(\widetilde{\mathbf{\Omega}}_{ij})/\beta}}_{=:F(\mathbf{\Omega};\widetilde{\mathbf{\Omega}})} + C \ ,$$

with equality at $\mathbf{\Omega} = \widetilde{\mathbf{\Omega}}$ (here C stands for the terms that are constant with respect to $\mathbf{\Omega}$). Note that $F(\mathbf{\Omega};\widetilde{\mathbf{\Omega}})$ is a convex function of $\mathbf{\Omega}$. Therefore, to find $\widehat{\mathbf{\Omega}}$,

1. Initialize $\mathbf{\Omega}_{[0]} = (\mathbf{0}_{p\times p}, \dots, \mathbf{0}_{p\times p})$ (or any other initial value).
2. For $t = 1, 2, \dots$, solve the convex optimization problem
$$\mathbf{\Omega}_{[t]} = \operatorname*{arg\,min}_{\mathbf{\Omega}\in\mathcal{S}_p(b)} F(\mathbf{\Omega}; \mathbf{\Omega}_{[t-1]}) \ . \tag{12}$$
3. Stop when some convergence criterion has been reached.

For optimizing (12), if $f(\cdot)$ is chosen to be the sparse group Lasso regularizer

$$f(\cdot) = \nu\|\cdot\|_1 + (1-\nu)\|\cdot\|_2 \ ,$$

then the step (12) is equivalent to a weighted group graphical Lasso problem [3], but with an additional constraint that $\mathbf{\Omega}_{[t]} \in \mathcal{S}_p(b)$; this constraint can be added to the ADMM algorithm for group graphical Lasso given in [3] with no additional computational cost.

If the objective function $F(\mathbf{\Omega})$ is convex over $\mathcal{S}_p(b)$—that is, if our choices of β, γ, and b satisfy the condition (11) of Theorem 1—then majorization-minimization is guaranteed to converge to a globally optimal solution $\widehat{\mathbf{\Omega}} \in \arg\min_{\mathbf{\Omega}\in\mathcal{S}_p(b)} F(\mathbf{\Omega})$ [16]. In practice, we may choose to remove the bound on the spectral norms, or equivalently, explore concavity of the penalty beyond what is allowed in the convexity condition (11), since lower values of β may perform better empirically.

3.3 Separation into Connected Components

For the graphical Lasso (1), Witten et al. [14] proved that the connected components of the solution $\widehat{\Omega}_{\mathsf{glasso}}$ can be identified in a preprocessing step that simply requires screening for sample correlations S_{ij} that exceed the penalty parameter value γ. This allows for significantly faster optimization of the graphical Lasso. Theorem 2 of Danaher et al. [3] extends this result to the group graphical Lasso setting, by screening for any $i < j$ such that

$$\sqrt{\sum_k \Big(n_k|S_{ij}^{(k)}| - \gamma\nu\Big)_+^2} > \gamma(1-\nu) \tag{13}$$

and then solving separate optimization problems for each resulting connected component. Their results prove that the combined solution yields a global minimizer $\widehat{\mathbf{\Omega}}_{\mathsf{GGL}}$ of the group graphical Lasso (2).

This type of block-wise optimization can be extended to the nonconvex log-shift penalty:

Theorem 2. *Consider any partition $\mathcal{A} = \{A_1, \dots, A_m\}$ of $[p]$ into disjoint sets. Suppose that*

$$-\gamma^{-1} \cdot \operatorname{diag}\{n_1, \dots, n_K\} \cdot \mathbf{S}_{ij} \in \partial f(\mathbf{0}) \textit{ for all } i \not\sim_{\mathcal{A}} j \tag{14}$$

where $\mathbf{S}_{ij} = (S_{ij}^{(1)}, \dots, S_{ij}^{(K)})$. If the conditions of Theorem 1 are satisfied, then there exists some $\widehat{\mathbf{\Omega}} \in \arg\min_{\mathbf{\Omega}\in\mathcal{S}_p(b)} F(\mathbf{\Omega})$ such that $\widehat{\Omega}_{ij}^{(k)} = 0$ for all k and all $i \not\sim_{\mathcal{A}} j$.

In particular, if $f(\cdot) = \nu\|\cdot\|_1 + (1-\nu)\|\cdot\|_2$, then condition (14) is equivalent to Danaher et al. [3]'s condition (13) for the group graphical Lasso. Although our penalty is nonconvex, near zero it is approximately equal to the group graphical Lasso penalty (or, more generally, $\beta\log\big(1 + f(\mathbf{\Omega}_{ij})/\beta\big) \approx f(\mathbf{\Omega}_{ij})$ when $\mathbf{\Omega}_{ij} \approx 0$). This allows us to extend the proof techniques of [3] to this nonconvex penalty setting. Theorem 2 is proved in the Supplementary Materials.

Based on this result, we now propose a faster algorithm for minimizing $F(\mathbf{\Omega})$.

1. Partition $[p]$ into sets $A_1, \dots, A_M$, the connected components of the adjacency matrix C, given by
$$C_{ij} = \mathbb{1}\left\{-\gamma^{-1} \cdot \operatorname{diag}\{n_1, \dots, n_K\} \cdot \mathbf{S}_{ij} \in \partial f(\mathbf{0})\right\} \ .$$
2. For $m = 1, \dots, M$, use majorization-minimization (Section 3.2) to solve the mth block,
$$\widehat{\mathbf{\Omega}}_m = \operatorname*{arg\,min}_{\mathbf{\Omega}\in\mathcal{S}_{p_m}(b)} F_m(\mathbf{\Omega}) \quad \text{where } p_m = |A_m| \text{ and}$$
$$F_m(\mathbf{\Omega}) = \sum_{k=1}^K -\frac{n_k}{2}\Big[\log\det(\Omega^{(k)}) - \langle \Omega^{(k)}, S_{A_m,A_m}^{(k)}\rangle\Big] \\ + \gamma \sum_{\substack{i<j \\ i,j\in A_m}} \beta\log(1 + f(\mathbf{\Omega}_{ij})/\beta) \ .$$

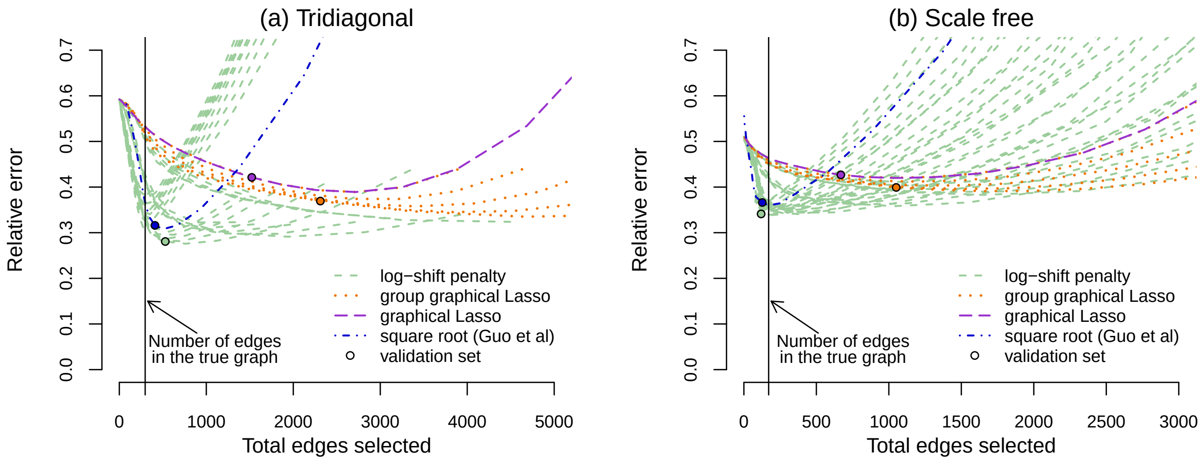

Figure 2: Experiment results for estimating precision matrices based on simulated data, plotting relative error in estimating $\mathbf{\Omega}$ versus the total number of edges selected is plotted. For each method, each line represents estimates with various values of γ while fixing other tuning parameters. (a) is for the tridiagnoal case and (b) for the blockwise scale free case. For each model, a held-out validation set was used to select tuning parameter value(s), highlighted in the plot. Plot is best viewed in color.

3. $\widehat{\mathbf{\Omega}}$ concatenates the blocks: $\widehat{\mathbf{\Omega}}_{A_m,A_m} = \widehat{\mathbf{\Omega}}_m$ for all m, and $\widehat{\mathbf{\Omega}}_{A_m,A_{m'}} = \mathbf{0}$ for all $m \neq m'$.

If the convexity condition (11) is satisfied, then Theorems 1 and 2 guarantee that the resulting solution $\widehat{\mathbf{\Omega}}$ is a global minimizer of $F(\mathbf{\Omega})$ over the set $\mathcal{S}_p(b)$.

4 EXPERIMENTS

4.1 Simulations

We implement our method on two sets of simulated data with different graph structures.

Tridiagonal graph data In our first example, we simulate $K = 3$ tridiagonal precision matrices of dimension $p = 100$, following the autoregressive (AR) process example in Fan et al. [4]. Specifically, for each k, the covariance matrix $\Sigma^{(k)}$ is defined as $\Sigma^{(k)}_{ij} = \exp(-|s^{(k)}_i - s^{(k)}_j|)$, where $0 = s^{(k)}_1 < s^{(k)}_2 < \cdots < s^{(k)}_p$ and $s^{(k)}_i - s^{(k)}_{i-1} \overset{\text{iid}}{\sim} \mathsf{Unif}(0.5, 1)$, $i = 2, \ldots, p$. Each precision matrix $\Omega^{(k)} = \left(\Sigma^{(k)}\right)^{-1}$ has the same support (they are each tridiagonal, due to the AR(1) covariance structure), but different values at the nonzero entries. For each k, we draw $n_k = 40$ i.i.d. samples from the distribution $\mathcal{N}\left(0, (\Omega^{(k)})^{-1}\right)$.

Scale-free graph data In the second example, we generate $K = 2$ graphs on $p = 100$ nodes. The first graph has 5 equally-sized scale-free subgraphs. The second graph shares the same structure in 4 subgraphs, but has no edge present in the remaining subgraph. The realization of the networks is depicted in Figure 3. A precision matrix is generated according to the first graph, with diagonal entries equal to 1, and off-diagonal entry values drawn from a uniform distribution on $[-0.4, -0.1] \cup [0.1, 0.4]$ when an edge is present, or otherwise 0. The precision matrix corresponding to the second graph is set to be identical to the first one, except that the off-diagonal entries are all set to zero for the subgraph where edges were removed.

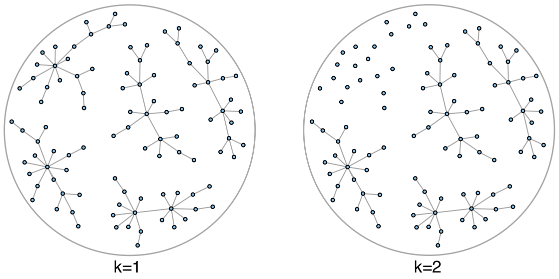

Figure 3: The realization of the blockwise scale-free graphs. The two graphs share the same structure in 4 of the 5 subgraphs. In the remaining subgraph, all edges are missing in the second graph.

Methods To implement our proposed method, we take $f(\cdot) = \nu\|\cdot\|_1 + (1-\nu)\|\cdot\|_2$ and minimize the objective function (8) for different values of tuning parameters (γ, β, ν). We also test the group graphical Lasso [3], the graphical Lasso [5], and Guo et al. [7]'s square-root method, for comparison.[2]

[2] Computations for simulations and for the real data experiment were performed in R [12] and used the `glasso` [6] and `JGL` [2] packages. Code for Guo et al. [7]'s method was obtained from the online supplementary material for [3], available at

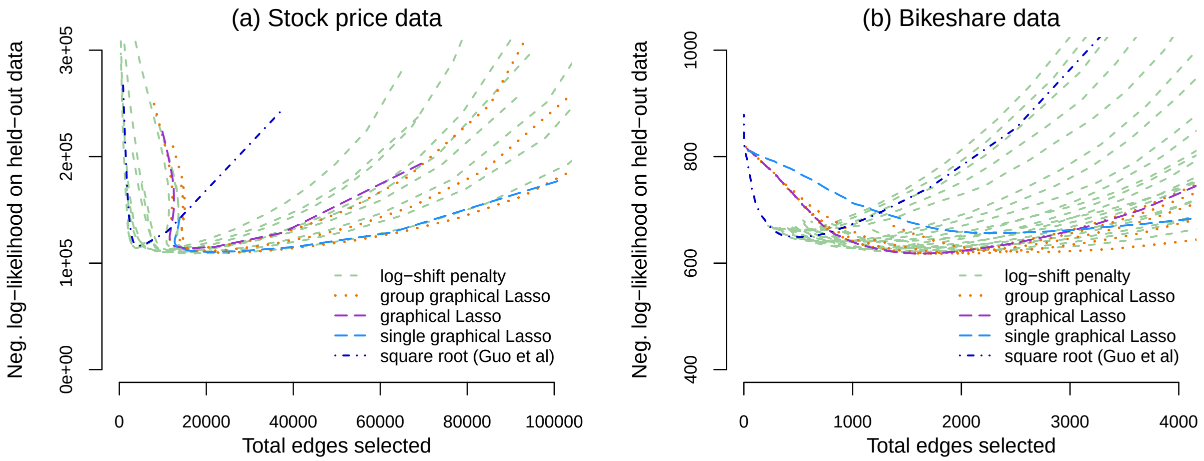

Figure 4: Experiment results for estimating precision matrices based on stock price data and bikeshare data. Negative log likelihood on held-out data is plotted versus the total number of edges selected. For each method, each line represents estimates with various values of γ while fixing other tuning parameters. Plot is best viewed in color.

Results Figure 2 displays results from the simulations. We plot the Frobenius norm of estimation error (normalized by the Frobenius norm of the true precision matrix) against the number of edges that are selected. Plots would be similar if the KL divergence between the estimator and the true precision matrix were plotted instead.

The graphs illustrate comparison on estimation and selection. In both experiments, among the methods considered, the log-shift method attains the lowest error in recovering $\mathbf{\Omega}$ and simultaneously selects an appropriately low number of edges, when tuning parameters are chosen judiciously (typically with a lower value of β, i.e. with high nonconvexity in the penalty). This suggests that the method is able to achieve relatively good estimation and selection with a single set of tuning parameters. As β increases, the performance of our method approaches that of the group graphical Lasso. In order to select appropriate tuning parameters for each method in a data-driven way, we generate a validation data set of same size and compute the log likelihoods of the validation data using the estimated precision matrices. For our model, the estimate selected by the validation score achieves the minimum error measure, and yields a total number of selected edges that is close to the number of true edges. We also notice that when tuning parameters are chosen such that the objective function is convex, the running time of our method is comparable to other convex methods. It becomes much slower when the objective function turns nonconvex.

4.2 Real Data

We next test our method on two sets of real data, stock price data and bikeshare data.

Stock price data We collect daily stock closing prices from Yahoo! Finance,[3] for $p = 432$ stocks that were consistently in the S&P 500 index from January 1, 2003 to December 31, 2012. Let $S_{i,j}$ be the closing price for stock j on day i, and $X_{i,j} = \log(S_{i,j}/S_{i-1,j})$ be the log return. We marginally transform the log returns of each stock to a normal distribution. Denoting the transformed data still as $X_{i,j}$, we treat the daily data $X_{i,\cdot} \in \mathbb{R}^p$ as independent observations, although they in fact form a time series. We divide the data into two time periods, one for before (2003–2007) and one for after (2009–2012) the 2008 financial crisis, and remove the data from 2008. The two sample sizes are $n_{\text{before}} = 1257$ and $n_{\text{after}} = 1005$. In the belief that the relationship between stocks might have changed during the financial crisis, we model the data as two GGMs (i.e. $K = 2$) with similar but non-identical precision matrices.

Capital bikeshare data We collect data from the Capital Bikeshare system,[4] a bike rental program in the D.C. area. It has a network of over 300 kiosk stations, where customers may rent a bike from a one location and return it to a different place on an as-needed basis, either as a "casual" user (paying for a single day) or a "registered" user (purchasing a membership). In the year 2012, there are $n = 366$ days and $p = 123$ high-activity stations. For $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, let $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ be the number of casual and registered rentals, respectively, initiated at station j on day i. After correcting for seasonal trend, we marginally transform each station's data to a normal distribution, and then model the data with two GGMs (i.e. $K = 2$, representing the casual rentals and the registered rentals), which we believe to have similar precision

http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1467-9868/homepage/76_2.htm.

[3] Data available at http://finance.yahoo.com

[4] Data available at https://www.capitalbikeshare.com/trip-history-data

matrix structure.

Methods For each of the data set, we select 20% of the data in each category as training data, and hold out the remaining 80% as the validation set. We implement our log-shift method with $K = 2$, and compare to the same existing methods as before. For an additional comparison, we also fit a single graphical Lasso to the combined data set (corresponding to the simple scenario that the two precision matrices have identical values in addition to identical edge structure).

Results To evaluate the results, we calculate the likelihood of the held-out data under the fitted models for each method. Results are displayed in Figure 4. For the stock price data, while the various methods' best scores are similar for the log-shift, group graphical Lasso, and single graphical Lasso, the log-shift method is able to attain this best validation score with a substantially smaller number of selected edges relative to the convex methods, demonstrating the benefit of the nonconvex penalty. For the bike-share data, both the log-shift method and group graphical lasso achieve the best validation score with a same number of edges selected. However, when moving from the optimal choice to a model with less edges, the convex methods suffer a large decline in performance on the validation set, while the log-shift method is able to maintain a nearly-optimal validation score.

5 DISCUSSION

In this paper, we introduce a family of nonconvex penalty functions, called the log-shift function, for estimating multiple related GGMs. It arises from a simple hierarchical model and generalizes existing methods for learning multiple GGMs, such as the group graphical Lasso [3]. Compared with methods that use a convex penalty function, the nonconvexity of the penalty function leads to less bias on strong signals and thus makes it possible to obtain good selection and estimation result at the same time. The log-shift penalty can also be applied to estimating other models, such as undirected graphical models with non-Gaussian distributions, time-varying GGMs, etc., which we leave to future work.

References

[1] Emmanuel J Candès, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

[2] Patrick Danaher. *JGL: Performs the Joint Graphical Lasso for sparse inverse covariance estimation on multiple classes*, 2013. URL http://CRAN.R-project.org/package=JGL. R package version 2.3.

[3] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical Lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.

[4] Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive Lasso and SCAD penalties. *The annals of applied statistics*, 3(2):521, 2009.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

[6] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. *glasso: Graphical Lasso- estimation of Gaussian graphical models*, 2011. URL http://CRAN.R-project.org/package=glasso. R package version 1.7.

[7] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[8] David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1): 30–37, 2004.

[9] Mladen Kolar, Le Song, Amr Ahmed, Eric P Xing, et al. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123, 2010.

[10] Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[11] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103 (482):681–686, 2008.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

[13] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[14] Daniela M Witten, Jerome H Friedman, and Noah Simon. New insights and faster computations for the graphical Lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

[15] Eleanor Wong, Suyash Awate, and P Thomas Fletcher. Adaptive sparsity in Gaussian graphical models. In *Proceedings of The 30th International*

Conference on Machine Learning, pages 311–319, 2013.

[16] CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[17] Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, (just-accepted):00–00, 2014.