Statistics and Its Interface Volume 14 (2021) 351–358

High-dimensional correlation matrix estimation for Gaussian data: a Bayesian perspective*

Chaojie Wang and Xiaodan Fan†

Gaussian covariance or precision matrix estimation is a classical problem in high-dimensional data analyses. For precision matrix estimation, the graphical lasso provides an efficient approach by optimizing the log-likelihood function with L_1-norm penalty. Inspired by the success of graphical lasso, researchers pursue analogous outcomes for covariance matrix estimation. However, it suffers from the difficulty of non-convex optimization and a degeneration problem when $p > n$ due to the singularity of the sample covariance matrix. In this paper, we fix the degeneration problem by adding an extra constraint on diagonal elements. From the Bayesian perspective, a grid-point gradient descent (GPGD) algorithm together with the block Gibbs sampler is developed to sample from the posterior distribution of the correlation matrix. The algorithm provides an effective approach to draw samples under the positive-definite constraint, and can explore the whole feasible region to attain the mode of the posterior distribution. Simulation studies and a real application demonstrate that our method is competitive with other existing methods in various settings.

Keywords and phrases: Correlation matrix estimation, Positive-definiteness, Non-convex optimization, Bayesian analysis.

1. BACKGROUND

Gaussian covariance or precision matrix estimation is a classical problem in high-dimensional data analyses. Under the Gaussian assumption, an estimator can be derived from the likelihood of observed data by the maximum likelihood estimation (MLE) criterion. In traditional multivariate Gaussian analyses, MLE of the covariance matrix $\mathbf{\Sigma}$ is proportional to the sample covariance matrix $\mathbf{S}$ when the sample size n is larger than the number of variables p. However, for high-dimensional cases, $\mathbf{S}$ is singular when $p > n$, leading the maximum of log-likelihood function to infinity [3]. Thus, it is necessary to introduce additional constraints to the objective function.

*This research is partially supported by two grants from the Research Grants Council of the Hong Kong SAR, China (CUHK14173817, CUHK14303819), one CUHK direct grant (4053357) and one UJS direct grant (5501190012).

†Corresponding author: xfan@cuhk.edu.hk.

[22] proposed the Gaussian graphical model, and illustrated the relationship between sparsity of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ and conditional independence of corresponding variables. As a consequence, it motivates many researchers to explore the sparse precision matrix estimation. [9] proposed an efficient algorithm called graphical lasso, which imposes L_1-norm penalty on entries of $\mathbf{\Omega}$. After that, many penalized approaches with different types of penalties on $\mathbf{\Omega}$ are proposed, such as adaptive graphical lasso [6], SCAD [7, 13], MCP [23], CLIME [4] and so on.

Graphical lasso achieved great success in precision matrix estimation. However, sparsity in a precision matrix usually does not result in a sparse covariance matrix. From the perspective of covariance matrix estimation, [2] followed the analogous idea to impose L_1 penalty on entries of $\mathbf{\Sigma}$ directly. The proposed estimator works well when n is large, but suffer from a degeneration problem when $p > n$ due to the singularity of the sample covariance matrix. [2] suggested to replace $\mathbf{S}$ with $\mathbf{S} + \epsilon\mathbf{I}$ for some positive ϵ when $p > n$, but the performance was still poor. Since the objective function in [2] is non-convex and may have multiple local maxima, it is hard to develop an effective algorithm to solve the optimization problem directly. Under this circumstance, Bayesian analysis provides a new perspective for this traditional estimation problem. It is well-known that L_1 penalty is equivalent to the Laplacian prior under the Bayesian framework. [21] and [12] proposed the Bayesian adaptive graphical lasso for precision matrix estimation. Recently, [10] considered to estimate a high-dimensional sparse precision matrix, in which adaptive shrinkage and sparsity are induced by a mixture of Laplace priors. With the similar ideas, [20] proposed the Bayesian covariance graphical lasso as a Bayesian version of the covariance graphical lasso proposed by [9]. They developed a coordinate descent algorithm to sample from the posterior distribution of the covariance matrix. However, they still suffered from the degeneration problem when $p > n$. For more details, [17] and [8] provided an overview of the works on covariance matrix estimation in the last decades.

The degeneration problem in [2] is resulted from the fact that the covariance matrix is not scale-invariant. [5] pointed that the ignorance of this prior information incurs p additional parameters in the diagonal entries of the estimate. Actually, it is a common setting to estimate variances with sample estimators. Then, according to the variance-

correlation decomposition, we only need to focus on the correlation matrix estimation. This paper contributes to the field in the following aspects: (a) we fix the degeneration problem by imposing an extra constraint on the diagonal entries of the estimate; (b) we develop a grid-point gradient descent algorithm together with the block Gibbs sampler to sample from the posterior distribution of the correlation matrix; (c) we show that the proposed algorithm provides an effective approach to draw samples under the positive-definite constraint, and can explore the whole feasible region to attain the mode of the posterior distribution; (d) our method is competitive with other existing methods in both simulations and real applications.

This paper is organized as follows: Section 2 introduces the estimator and the algorithms in details; Section 3 performs extensive simulation studies to compare our method with existing approaches; Section 4 provides a real application; Section 5 concludes the paper.

2. METHODS

The observation $\mathbf{X}_{n\times p}$ is a data set comprising n samples associated with p variables. Let $\boldsymbol{X}_i$, $i = 1, \cdots, n$, be i.i.d samples and follow a p-dimensional normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where the mean vector is assumed to be zero's without loss of generality and $\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p}$ is an unknown covariance matrix. Data following a general p-dimensional normal distribution can be reduced to our problem by centralization. Then the log-likelihood function of observed data is

$$\ell(\boldsymbol{\Sigma}) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log\det\boldsymbol{\Sigma} - \frac{n}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}),$$

where $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ is the sample covariance matrix given the zero-mean assumption. In this paper, we are interested in estimating the covariance matrix with fixed p and finite sample size n when $p > n$.

[2] proposed the following penalized log-likelihood method with L_1 penalty on entries of $\boldsymbol{\Sigma}$ directly:

$$\underset{\boldsymbol{\Sigma}\succ 0}{\text{Minimize}} \quad \log\det\boldsymbol{\Sigma} + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}) + \rho||\boldsymbol{\Sigma}||_1, \tag{1}$$

where ρ is a nonnegative tuning parameter and $||\boldsymbol{\Sigma}||_1 = \sum_{i\neq j}|\sigma_{ij}|$. They developed a majorization-minimization approach to solve Equation 1. However, it meets the degeneration problem when $p > n$. Actually, unlike the graphical lasso, imposing L_1 penalty on $\boldsymbol{\Sigma}$ can not ensure the existence of optima in Equation 1. Since $\mathbf{S}$ is not of full rank when $p > n$, there exists an eigenvector $\boldsymbol{v} \neq \mathbf{0}$ such that $\mathbf{S}\boldsymbol{v} = \mathbf{0}$. Let $\mathbf{U} = \sum_{j=1}^{p-1}\boldsymbol{u}_j\boldsymbol{u}_j'$ be an orthogonal matrix where $\boldsymbol{u}_j$'s are eigenvectors satisfying $\boldsymbol{u}_j \perp \boldsymbol{v}$ for $j = 1, \cdots, p-1$. An estimator can be constructed as

$$\hat{\boldsymbol{\Sigma}} = \alpha\boldsymbol{v}\boldsymbol{v}' + \mathbf{U}.$$

Then we have

$$\log\det\hat{\boldsymbol{\Sigma}} = \log\left(\alpha\prod_{j=1}^{p-1}1\right) = \log\alpha,$$
$$\mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S}) = \mathrm{tr}\left(\left(\frac{1}{\alpha}\boldsymbol{v}\boldsymbol{v}' + \mathbf{U}\right)\mathbf{S}\right) = \mathrm{tr}(\mathbf{U}\mathbf{S}).$$

Thus for Equation 1,

$$\log\alpha + \mathrm{tr}(\mathbf{U}\mathbf{S}) + \rho||\alpha\boldsymbol{v}\boldsymbol{v}' + \mathbf{U}||_1 \to -\infty$$

as $\alpha \to 0$. The objective function goes into infinity when $p > n$.

In nature, the degeneration problem is resulted from the fact that the covariance matrix is not scale-invariant. [5] pointed that the ignorance of this prior information incurs p additional parameters in the diagonal entries of the estimate. Based on this idea, we propose a modified estimator by adding an extra constraint to Equation 1. The objective function can be expressed as follows:

$$\begin{aligned}\underset{\boldsymbol{\Sigma}\succ 0}{\text{Minimize}} \quad & \log\det\boldsymbol{\Sigma} + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}) + \rho||\boldsymbol{\Sigma}||_1 \\ s.t. \quad & \sigma_{jj} = s_{jj}, \quad j = 1, \cdots, p.\end{aligned} \tag{2}$$

Actually, according to the variance-covariance decomposition, $\boldsymbol{\Sigma} = \mathbf{D}\boldsymbol{\Gamma}\mathbf{D}$, where $\mathbf{D}$ is the diagonal matrix of standard deviations and $\boldsymbol{\Gamma}$ is the correlation matrix with diagonal elements equal to 1. Thus, we may estimate $\mathbf{D}$ and $\boldsymbol{\Gamma}$ separately [1]. If $\mathbf{D}$ is estimated by the sample variance as common choices, i.e., $\hat{\mathbf{D}} = \mathrm{diag}(\mathbf{S})^{\frac{1}{2}}$, then we can focus on the estimation of the correlation matrix $\boldsymbol{\Gamma} = (\gamma_{ij})_{p\times p}$. By replacing the sample covariance matrix $\mathbf{S}$ with the sample correlation matrix $\mathbf{R}$, the problem can be rewritten as follows:

$$\begin{aligned}\underset{\boldsymbol{\Gamma}\succ 0}{\text{Minimize}} \quad & \log\det\boldsymbol{\Gamma} + \mathrm{tr}(\boldsymbol{\Gamma}^{-1}\mathbf{R}) + \rho||\boldsymbol{\Gamma}||_1 \\ s.t. \quad & \gamma_{jj} = 1, \quad j = 1, \cdots, p,\end{aligned} \tag{3}$$

where $\mathbf{R} = \hat{\mathbf{D}}^{-1}\mathbf{S}\hat{\mathbf{D}}^{-1}$. Since $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$, we have $\mathbf{R} = \frac{1}{n}\mathbf{Y}'\mathbf{Y}$, where $\mathbf{Y} = \mathbf{X}\hat{\mathbf{D}}^{-1}$ is the data after normalization. The following sections focus on the estimation of $\boldsymbol{\Gamma}$ in Equation 3 for convenience.

2.1 Bayesian perspectives

Note that the objective function in Equation 3 is non-convex and may have multiple local maxima. Traditional algorithms are hard to solve this optimization problem. Inspired by the ideas in [20], we propose a Bayeisan algorithm to sample from the posterior distribution of $\boldsymbol{\Gamma}$. It is well-known that L_1 penalty is equivalent to Laplacian prior in Bayesian perspectives. Thus, the optimum of Equation 3 can

be pursued by solving the following Bayesian model:

$$p(\mathbf{Y}|\mathbf{\Gamma}) \propto |\mathbf{\Gamma}|^{-n/2} \exp\left(-\mathrm{tr}\left(\frac{n}{2}\mathbf{\Gamma}^{-1}\mathbf{R}\right)\right),$$

$$p(\mathbf{\Gamma}|\lambda) \propto \prod_{i<j} h(\gamma_{ij}|\lambda) \cdot \prod_{j=1}^{p} 1_{\{\gamma_{jj}=1\}} \cdot 1_{\mathbf{\Gamma}\succ 0},$$

where $h(x|\lambda)$ represents the double exponential (laplace) density function with the form $\lambda/2\exp(-\lambda|x|)$, and $\mathbf{\Gamma} \succ 0$ denotes that $\mathbf{\Gamma}$ is positive-definite. Then the posterior distribution can be written as

$$\begin{aligned} p(\mathbf{\Gamma}|\mathbf{Y},\lambda) \propto & |\mathbf{\Gamma}|^{-n/2} \exp\left(-\mathrm{tr}\left(\frac{n}{2}\mathbf{\Gamma}^{-1}\mathbf{R}\right)\right) \\ & \cdot \prod_{i<j} h(\gamma_{ij}|\lambda) \cdot \prod_{j=1}^{p} 1_{\{\gamma_{jj}=1\}} \cdot 1_{\mathbf{\Gamma}\succ 0}. \end{aligned} \tag{4}$$

The mode of posterior distribution of $\mathbf{\Gamma}$ is equivalent to the estimate in Equation 3 with $\rho = \lambda/n$. Thus the problem is transformed into the search for the maximum a posterior of Equation 4.

2.2 Posterior inference

We adopt the idea of block Gibbs sampler to update $\mathbf{\Gamma}$ one column and one row at a time while holding all of the rest elements in $\mathbf{\Gamma}$ fixed. Without loss of generality, we focus on the last column and last row. Partition $\mathbf{\Gamma}$ and $\mathbf{R}$ as follows:

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_{11} & \boldsymbol{\gamma}_{12} \\ \boldsymbol{\gamma}_{12}' & \gamma_{22} \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{r}_{12} \\ \mathbf{r}_{12}' & 1 \end{pmatrix},$$

where $\mathbf{\Gamma}_{11}$ and $\mathbf{R}_{11}$ are the submatrix of the first $p-1$ rows and columns.

Conditioning on $\mathbf{\Gamma}_{11}$ and $\gamma_{22} = 1$,

$$\begin{aligned} & p(\boldsymbol{\gamma}_{12}|\mathbf{\Gamma}_{11}, \gamma_{22}=1, \mathbf{Y}, \lambda) \\ & \propto |\mathbf{\Gamma}|^{-n/2} \exp\left(-\mathrm{tr}\left(\frac{n}{2}\mathbf{\Gamma}^{-1}\mathbf{R}\right)\right) \cdot h(\boldsymbol{\gamma}_{12}|\lambda) \cdot 1_{\mathbf{\Gamma}\succ 0} \\ & \propto |\mathbf{\Gamma}|^{-n/2} \exp\left(-\mathrm{tr}\left(\frac{n}{2}\mathbf{\Gamma}^{-1}\mathbf{R}\right)\right) \cdot \exp\left(-\lambda \sum_{j=1}^{p-1} |\gamma_{pj}|\right) \cdot 1_{\mathbf{\Gamma}\succ 0}. \end{aligned}$$

In the iterative procedure, $\mathbf{\Gamma} \succ 0$ leads to $\mathbf{\Gamma}_{11} \succ 0$. Given $\gamma_{22} = 1$, the next proposed $\tilde{\mathbf{\Gamma}}$ is positive-definite if and only if the proposed $\tilde{\boldsymbol{\gamma}}_{12}$ satisfies that $1 - \tilde{\boldsymbol{\gamma}}_{12}'\mathbf{\Gamma}_{11}^{-1}\tilde{\boldsymbol{\gamma}}_{12} > 0$. Let $H = 1 - \tilde{\boldsymbol{\gamma}}_{12}'\mathbf{\Gamma}_{11}^{-1}\tilde{\boldsymbol{\gamma}}_{12}$, then the posterior distribution of $\boldsymbol{\gamma}_{12}$ can be written as

$$\begin{aligned} & p(\boldsymbol{\gamma}_{12}|\mathbf{\Gamma}_{11}, \gamma_{22}=1, \mathbf{Y}, \lambda) \\ & \propto |\mathbf{\Gamma}|^{-n/2} \exp\left(-\mathrm{tr}\left(\frac{n}{2}\mathbf{\Gamma}^{-1}\mathbf{R}\right) - \lambda \sum_{j=1}^{p-1} |\gamma_{pj}|\right) \cdot 1_{H>0}. \end{aligned} \tag{5}$$

A grid-point gradient descent algorithm is developed to sample from the conditional posterior distribution in Equation 5. The steps are summarized in Algorithm 1. Basically, we propose a new sample $\tilde{\boldsymbol{\gamma}}_{12}$ from the grid points around the original sample $\boldsymbol{\gamma}_{12}$. The proposal probability of the grid points, which do not satisfy the positive-definiteness, are set to be 0. An acceptance ratio is designed to ensure the Markov chain is reversible.

Algorithm 1 Grid-Point Gradient Descent Algorithm (GPGD)

1: Sample a random tiny step $\epsilon \sim \mathrm{Uniform}(0, 0.05)$.
2: Find all $2(p-1)$ grid points around $\boldsymbol{\gamma}_{12}$ at radius ϵ, i.e.,
3: $\quad \boldsymbol{\gamma}_{12}^{(h)} = \boldsymbol{\gamma}_{12} + \epsilon \boldsymbol{I}_h,\ h = 1, \cdots, p-1,$
4: $\quad \boldsymbol{\gamma}_{12}^{(h)} = \boldsymbol{\gamma}_{12} - \epsilon \boldsymbol{I}_{h-p+1},\ h = p, \cdots, 2(p-1),$
5: where $\boldsymbol{I}_h$ is a $p-1$-dimensional vector with the h-th element equal to 1 and others equal to 0.
6: Calculate the posterior probability $p(\boldsymbol{\gamma}_{12}|\mathbf{\Gamma}_{11}, \gamma_{22}=1, \mathbf{Y}, \lambda)$ for every $\boldsymbol{\gamma}_{12}^{(h)}$.
7: Propose the new $\tilde{\boldsymbol{\gamma}}_{12}$ with maximal posterior.
8: Calculate the posterior probability $p(\tilde{\mathbf{\Gamma}}|\mathbf{Y}, \lambda)$ based on $\tilde{\boldsymbol{\gamma}}_{12}$.
9: Accept the new proposal with probability $\min(1, r)$ where
10: $\quad r = \exp\{\log \frac{p(\tilde{\mathbf{\Gamma}}|\mathbf{Y},\lambda)}{p(\mathbf{\Gamma}|\mathbf{Y},\lambda)} \cdot Q_t\}$

Algorithm 1 draws samples of $\boldsymbol{\gamma}_{12}$ from the posterior distribution of Equation 5. The entire procedures of sampling from the posterior of $\mathbf{\Gamma}$ are concluded in Algorithm 2. Algorithm 2 implements a random scan Gibbs sampler to update the parameters column by column [15]. The initial value of $\mathbf{\Gamma}$ must be a symmetric positive-definite matrix with diagonal elements equal to 1. The default choice is $(1-\pi)\mathrm{diag}(\mathbf{R}) + \pi\mathbf{R}$ where $\pi \in [0, 1)$. The iteration steps are repeated until convergence.

Algorithm 2 Block Gibbs Algorithm

1: Initial $\mathbf{\Gamma}^{(0)}$.
2: **for** t-th iteration **do**
3: $\quad$ Sample $j \in \{1, \cdots, p\}$ uniformly
4: $\quad$ Sample $\tilde{\boldsymbol{\gamma}}_{12}$ instead of $\boldsymbol{\gamma}_{12}$ in the j-th column and row of $\mathbf{\Gamma}^{(t-1)}$ by implementing Algorithm 1.
5: **end for**

In Algorithm 1, Q_t is a tuning parameter to accelerates the speed of convergence, where $t = 1, 2, \cdots, T$ denotes the t-th iteration in Algorithm 2. It is set as a large value initially and then decays to 1. Specifically, the total iterations $T = 1500p$ is used for our simulation studies and the real application in this paper. Then, we consider Q_t a linearly decreasing sequence from 10 to 1 in the first $1000p$ iterations and remains 1 in the last $500p$ iterations. Thus, the sampling procedure will be fast at the beginning and converge to the posterior distribution finally.

2.3 Theoretical results

This section shows that Algorithm 2 is reversible and can explore the whole feasible space to attain the mode of the posterior distribution. For convenience, a concept called *positive-definite path* is defined as follows:

Definition 1 (Positive-definite Path)**.** *For two positive-definite matrices with the same dimension,* $\mathbf{A}$ *and* $\mathbf{B}$*, if there exists a continuous evolution from* $\mathbf{A}$ *to* $\mathbf{B}$ *and all states in the evolution path remain positive-definite, this evolution is called a positive-definite path and denoted by* $\mathbf{A} \to \mathbf{B}$*.*

Then we have the following theorem:

Theorem 1. *For any two positive-definite matrices* $\mathbf{A}_{p\times p}$ *and* $\mathbf{B}_{p\times p}$*, there exists a positive-definite path that* $\mathbf{A}_{p\times p} \to \mathbf{B}_{p\times p}$*.*

See Appendix A for detailed proofs of Lemmas, Corollaries and Theorems.

Corollary 1. *For any two positive-definite matrices* $\mathbf{A}_{p\times p}$ *and* $\mathbf{B}_{p\times p}$ *with the same diagonal elements, i.e.,* $a_{jj} = b_{jj}$*,* $j = 1, \cdots, p$*, there exists a positive-definite path that* $\mathbf{A}_{p\times p} \to \mathbf{B}_{p\times p}$ *without changing the diagonal elements.*

According to Corollary 1, for any symmetric positive-definite matrix with diagonal elements equal to 1, there exists a positive-definite path from the initial $\mathbf{\Gamma}^{(0)}$ to it without changing the diagonal elements. Then we show that this positive-definite path can be achieved by Algorithm 2.

Lemma 1. *Define*

$$\mathbf{J}_h = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I}_h \\ \boldsymbol{I}_h' & 0 \end{pmatrix}, \quad h = 1, \cdots, p-1,$$

where $\boldsymbol{I}_h$ *is a* $(p-1)$*-dimensional vector with the* h*-th element equal to 1 and others equal to 0. If* $\mathbf{\Gamma}$ *is a symmetric positive-definite matrix, then* $\mathbf{\Gamma} \pm \epsilon \mathbf{J}_h$ *is positive-definite for any* h *as* $\epsilon \to 0$*.*

Theorem 2. *Any symmetric positive-definite matrix with diagonal elements equal to 1 can be attained by Algorithm 2.*

Algorithm 2 provides an effective approach to draw samples under the positive-definite constraint. All proposed samples in the procedure remain positive-definite. Besides, it is reversible and can explore the whole feasible space to attain the mode of posterior distribution. Thus, it can solve the non-convex optimization problem in Equation 3.

3. SIMULATIONS

In this section, simulation studies are conducted to compare the performance of our method with other existing approaches. The compared approaches include the majorize-minimize (MM) algorithm [2], the classical graphical lasso (GLasso) [9], the log-likelihood with SCAD penalty on the precision matrix (SCAD) [6], the log-likelihood with MCP penalty on the precision matrix (MCP) [23], the positive definite sparse covariance estimators (PDSCE) under the Frobenius norm [18], and the shrinkage estimator [14]. Two criteria are used to evaluate the performance of estimators: penalized negative log-likelihood ℓ and Matthews correlation coefficient (MCC) [6].

Penalized negative log-likelihood ℓ, which measures the fitness of estimates to the observed data with penalty, is defined as follows:

$$\ell(\mathbf{\Sigma}) = \log\det \mathbf{\Sigma} + \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S}) + \rho||\mathbf{\Sigma}||_1.$$

Here $\ell(\mathbf{\Sigma})$ not only measures the performance of the correlation matrix estimator, but also considers the estimations of variances. The smaller ℓ, the better fitness to the observed data. So it is fair to compare our GPGD with other methods that focus on covariance matrix estimation.

MCC, which measures the accuracy of sparsity specification, i.e., how many zero correlations are classified as non-zeros and how many non-zero correlations are classified as zeros, is defined as follows:

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

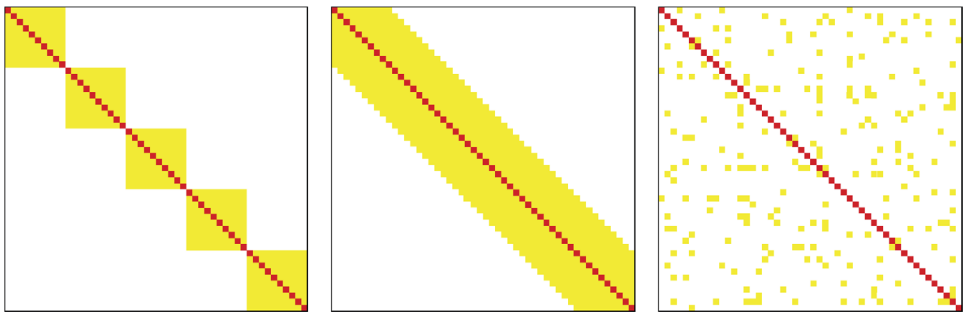
where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives, respectively. Here the larger MCC means more accurate specification of sparsity.

In the following simulation studies, two types of common-used sparse structure are considered. The synthesized data are generated from a given data model with a known correlation matrix. Considering the uncertainty of Monte Carlo simulations, the experiments, including the data synthesis, are repeated for 100 times independently in each setting. The means and standard errors of ℓ and MCC are reported for comparison. See the supplementary materials for the detailed R codes, http://intlpress.com/site/pub/files/_supp/sii/2021/0014/0003/SII-2021-0014-0003-s003.zip.

3.1 Case 1: block-diagonal models

In this case, data are generated with a block-diagonal covariance matrix referring to [2] and [18]. The true covariance matrix is set as $\mathbf{\Sigma} = \mathrm{diag}(\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_5)$, where $\mathbf{\Sigma}_k$, $k = 1, \cdots, 5$, are general dense matrices with dimensions $(p/5)\times(p/5)$. The off-diagonal elements of $\mathbf{\Sigma}_k = (\sigma_{ij}^{(k)})$ are sampled independently from the uniform distribution $(-1, 1)$, i.e., $\sigma_{ij}^{(k)} = \sigma_{ji}^{(k)} \sim \mathrm{Uniform}(-1, 1)$ for $i \neq j$. The diagonal elements of $\mathbf{\Sigma}_k$ are set as a constant so that the resulting matrix $\mathbf{\Sigma}$ has condition number equal to p as in [19]. The covariance matrix $\mathbf{\Sigma}$ corresponds to a graph with 5 disconnected cliques of size $p/5$. See Figure 1 (a) for detailed structures. Note that the inverse of block-diagonal matrices are also block diagonal. Thus, the approaches designed for sparse precision matrices, such as graphical lasso, could be compared in the same model fairly. The simulation results are presented in Table 1.

Table 1 demonstrates that our GPGD estimator outperforms other approaches on penalized negative log-likelihood ℓ significantly. It means the GPGD estimator has the best fitness to the observed data. For MCC, our GPGD is close



(a) Block-diagonal models (b) Banded models (c) Randomly sparse models

Figure 1. The structures of each model. The red areas denote the diagonal elements. The yellow areas denote the non-zero off-diagonal elements. The white areas denote the diagonal elements with zeros.

Table 1. The means and standard errors of two criteria across 100 independent experiments for block-diagonal cases, where the tuning parameter $\rho = 0.1$

	$p = 50$, $n = 25$		$p = 100$, $n = 50$	
	ℓ	MCC	ℓ	MCC
GPGD	**114.4780(9.1870)**	**0.6024(0.0031)**	**350.1483(16.9879)**	0.5860(0.0073)
MM	151.3274(12.9590)	0.4513(0.0187)	532.8467(30.2572)	0.4915(0.0082)
GLasso	189.9787(15.8022)	0.1786(0.0325)	718.3091(41.5621)	0.1752(0.0171)
SCAD	187.2170(16.0069)	0.1765(0.0333)	718.3265(42.0373)	0.1757(0.0152)
MCP	184.3702(15.8384)	0.1776(0.0343)	720.2351(41.8870)	0.1765(0.0168)
PDSCE	152.4892(12.1424)	0.5996(0.0057)	462.3246(21.9841)	**0.6038(0.0012)**
Shrinkage	177.0516(15.2729)	0.1930(0.0358)	661.5784(37.4026)	0.1856(0.0166)

Table 2. The means and standard errors of two criteria across 100 independent experiments for banded cases, where the tuning parameter $\rho = 0.1$

	$p = 50$, $n = 25$		$p = 100$, $n = 50$	
	ℓ	MCC	ℓ	MCC
GPGD	**118.9781(9.5368)**	**0.5860(0.0032)**	**273.3007(13.7905)**	0.7768(0.0047)
MM	158.3069(13.7026)	0.4388(0.0205)	392.8170(23.4201)	0.6852(0.0053)
GLasso	197.8926(17.1376)	0.1752(0.0295)	520.8988(31.4136)	0.3376(0.0213)
SCAD	195.0956(16.8915)	0.1703(0.0345)	521.1440(31.4554)	0.3350(0.0213)
MCP	192.1365(16.7525)	0.1712(0.0371)	523.6765(31.3181)	0.3304(0.0221)
PDSCE	158.3995(12.8966)	0.5841(0.0057)	352.8728(18.2103)	**0.7893(0.0007)**
Shrinkage	185.0087(16.3894)	0.1721(0.0318)	482.9550(28.9142)	0.3450(0.0193)

to PDSCE and both of them have significant advantages over others. GLasso, SCAD and MCP methods have been shown to perform well in estimating sparse precision matrices, but Table 1 demonstrates that they fail to generate sparse covariance matrix estimators. All of them have poor performance in MCC.

3.2 Case 2: banded models

In this cases, data are generated with a banded covariance matrix. The banded model is also a common-used setting of covariance matrices and considered in [2] and [18]. The bandwidth of the true covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ is set as 5, i.e., $\sigma_{ij} = 0$ if $|i-j| > 5$; $\sigma_{ij} = \sigma_{ji}$ are sampled independently from the uniform distribution $(-1, 1)$ if $0 < |i-j| \leq 5$; The diagonal elements of $\mathbf{\Sigma}$ are set as a constant so that it has condition number equal to p as in [19]. See Figure 1 (b) for detailed structures. The simulation results are presented in Table 2. Again, the results show that our method is competitive with others.

3.3 Case 3: randomly sparse models

For more general cases, we consider a randomly sparse covariance matrix referring to [2]. Each non-diagonal element $\sigma_{ij} = \sigma_{ji}$ is sampled independently from the uniform distribution $(-1, 1)$, and then is reset as 0 with probability 90%. So we can obtain a randomly sparse covariance matrix. The diagonal elements of $\mathbf{\Sigma}$ are set as a constant so that it has condition number equal to p. See Figure 1 (c) for details.

Table 3. The means and standard errors of two criteria across 100 independent experiments for randomly sparse cases, where the tuning parameter $\rho = 0.1$

	$p = 50$, $n = 25$		$p = 100$, $n = 50$	
	ℓ	MCC	ℓ	MCC
GPGD	**95.7640(7.7047)**	**0.7662(0.0160)**	**284.4719(12.3516)**	0.7730(0.0102)
MM	123.0414(11.3930)	0.6182(0.0218)	411.9183(21.6392)	0.6799(0.0126)
GLasso	155.1158(13.8076)	0.3129(0.0399)	545.5620(29.2087)	0.3292(0.0219)
SCAD	152.7447(13.9245)	0.3058(0.0376)	545.7803(29.1427)	0.3239(0.0236)
MCP	150.1693(13.7904)	0.3005(0.0396)	548.2684(29.0411)	0.3154(0.0224)
PDSCE	125.4417(10.7832)	0.7613(0.0165)	366.7575(17.0446)	**0.7853(0.0086)**
Shrinkage	143.1542(13.5374)	0.3151(0.0383)	505.7313(26.8765)	0.3303(0.0242)

The simulation results are presented in Table 3. They draw similar conclusions with that our method is competitive with others.

From above three data models, we can conclude that our GPGD estimator has significant advantages in fitting observed data over other competitors. It also provides a good sparse property in estimating sparse covariance matrices. Under the similar framework with majorize-minimize algorithm in [2], the GPGD algorithm provides a better algorithm to attain the optima.

4. APPLICATION ON PORTFOLIO OPTIMIZATION

In this section, we consider a real application on portfolio optimization. [16] proposed the famous modern portfolio theory (MPT), which models a portfolio of assets with means and variances of return sequences. Here we work on an improved version of MPT called minimum variance portfolio (MVP), proposed by [11].

Given $X_{n\times p}$ is the return dataset of p assets, which is assumed to follow a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, MPV minimizes the variance of portfolio under a constraint of the lowest required return μ_0, i.e.,

$$\begin{aligned} \underset{\mathbf{w}}{\text{Minimize}} \quad & \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \\ s.t. \quad & \mathbf{w}'\boldsymbol{\mu} \geq \mu_0, \\ & w_j \geq 0, \quad j = 1, 2, \cdots, p, \\ & \sum_{j=1}^{p} w_j = 1, \end{aligned} \tag{6}$$

where $\mathbf{w} = (w_1, \cdots, w_p)'$ is a weight vector of portfolio. The solution of Equation 6 is a simple quadratic optimization with linear constraints. The difficulty is the estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Here we estimate $\boldsymbol{\mu}$ by sample mean of observed data. And $\boldsymbol{\Sigma}$ are considered to estimate in different ways, including our GPGD and compared approaches.

We obtain real financial data from Wind Financial Terminal (WTF). The dataset is the monthly returns of constituent stocks of China Securities Index 100 (CSI100), which is comprised of the 100 largest companies by market capitalization in Chinese stock market. The period of data is considered from 2015 to 2019, totally 60 observations. After deleting the stocks with missing observations (suspended more than one month), it remains 63 stocks ($p = 63$) in the pool. The length of train data used to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is 36 months ($n = 36$). Then, according to the optimal portfolio weight solved by Equation 6, we construct the portfolio and calculate the standard error of portfolio returns in the following next 12 months as test data. The smaller standard error means lower risks of portfolio, which is considered better in the framework of MVP. Table 4 presents the standard errors in test data for various approaches to estimate covariance matrix $\boldsymbol{\Sigma}$.

Table 4. The standard error (%) of portfolio in the test data

End Month	Oct 2019	Nov 2019	Dec 2019
GPGD	4.90	5.09	4.23
MM	5.35	5.57	4.59
GLasso	6.49	6.93	5.51
SCAD	6.25	6.47	4.94
MCP	6.29	6.51	4.94
PDSCE	5.89	6.15	4.95
Shrinkage	6.28	6.54	5.22

Here "End Month" means the last month in the test data. For example, the end month Dec 2019 represents the test dataset from Jan 2019 to Dec 2019 (12 months) and the train dataset from Jan 2016 to Dec 2018 (36 months). To demonstrate the sensitivity of estimations, we also consider the moving windows one month ahead for Nov 2019 and Oct 2019. All of them show that our GPGD estimator provides a portfolio with the lowest risk and thus has the best performance.

5. SUMMARY

Gaussian covariance matrix estimation is a basic and classical problem in high-dimensional areas. There have been many works on this problem. To make full use of the Gaussian assumption, optimization of penalized negative log-likelihood functions becomes the mainstream. The difference among various methods concentrates on the forms of

penalty terms and the objects of penalty: covariance or precision matrices. From the perspective of covariance matrices, [2] imposed L_1-norm penalty on entries of $\mathbf{\Sigma}$ directly to pursue the sparsity in covariance matrices. Unfortunately, this estimator leads to the degeneration problem when $p > n$. Due to the non-convexity of objective function, it is hard to design an efficient algorithm to solve the optimization directly.

In this paper, we gave our new insights of this classical problem from a Bayesian perspective. To avoid the degeneration problem in [2], an extra constraint is introduced by fixing the diagonal entries of the estimate to be unity. We developed an efficient gradient descent algorithm to sample from the posterior distribution, which may be non-convex and have multiple local maxima. Theoretically, we show that this algorithm would remain positive-definite in the whole sampling procedure. Through simulation studies and a real application, we compare the performance of our algorithm with other existing approaches. These empirical evidences demonstrate that our estimator is competitive with others in different indices.

Note that although we only compare the point estimate with existing methods, our Bayesian approach can actually provide the full posterior distribution. The Bayesian perspective also helps to design an efficient algorithm which guarantees the positive-definiteness along the path. Some theoretical aspects can be explored in future research. For example, the consistency of our estimator and the convergence rate of our algorithm are valuable things to explore.

APPENDIX A. THEORETICAL PROOFS

A.1 Theorem 1

Let $\mathbf{\Delta} = \mathbf{B} - \mathbf{A}$. Since $\mathbf{A}$ and $\mathbf{B} = \mathbf{A} + \mathbf{\Delta}$ are positive-definite, then for any $0 < \eta < 1$, $\mathbf{A} + \eta\mathbf{\Delta} = \eta(\mathbf{A} + \mathbf{\Delta}) + (1 - \eta)\mathbf{A}$ is positive-definite. Thus there exists a positive-definite path that $\mathbf{A} \xrightarrow{+\eta\mathbf{\Delta}} \mathbf{B}$ with $\eta : 0 \to 1$.

A.2 Corollary 1

If $a_{jj} = b_{jj}$ for $j = 1, \cdots, p$, then the diagonal elements of $\mathbf{\Delta} = \mathbf{B} - \mathbf{A}$ are all 0. Thus, for any $0 < \eta < 1$, the diagonal elements of $\mathbf{A} + \eta\mathbf{\Delta}$ do not change in the evolution.

A.3 Lemma 1

Here we only prove the case that $\mathbf{\Gamma} + \epsilon\mathbf{J}_h$ is positive-definite. The other case $\mathbf{\Gamma} - \epsilon\mathbf{J}_h$ is similar.

Partition $\mathbf{\Gamma}$ as follows:

$$\tilde{\mathbf{\Gamma}} = \mathbf{\Gamma} + \epsilon\mathbf{J}_h = \begin{pmatrix} \mathbf{\Gamma}_{11} & \boldsymbol{\gamma}_{12} + \epsilon\boldsymbol{I}_h \\ \boldsymbol{\gamma}'_{12} + \epsilon\boldsymbol{I}'_h & \gamma_{22} \end{pmatrix}.$$

Then apply the block matrix determinant

$$\begin{aligned} \det(\tilde{\mathbf{\Gamma}}) &= \det(\mathbf{\Gamma}_{11})(\gamma_{22} - (\boldsymbol{\gamma}'_{12} + \epsilon\boldsymbol{I}'_h)\mathbf{\Gamma}_{11}^{-1}(\boldsymbol{\gamma}_{12} + \epsilon\boldsymbol{I}_h)) \\ &= \det(\mathbf{\Gamma}_{11})\big[(\gamma_{22} - \boldsymbol{\gamma}'_{12}\mathbf{\Gamma}_{11}^{-1}\boldsymbol{\gamma}_{12}) - 2\epsilon\boldsymbol{\gamma}_{12}\mathbf{\Gamma}_{11}^{-1}\boldsymbol{I}'_h \\ &\qquad - \epsilon^2\boldsymbol{I}'_h\mathbf{\Gamma}_{11}^{-11}\boldsymbol{I}_h\big]. \end{aligned}$$

Because $\mathbf{\Gamma}$ is positive-definite, both of the terms $\det(\mathbf{\Gamma}_{11})$ and $(\gamma_{22} - \boldsymbol{\gamma}'_{12}\mathbf{\Gamma}_{11}^{-1}\boldsymbol{\gamma}_{12})$ are positive. Thus, there must exist a small $\epsilon > 0$ that $[(\gamma_{22} - \boldsymbol{\gamma}'_{12}\mathbf{\Gamma}_{11}^{-1}\boldsymbol{\gamma}_{12}) - 2\epsilon\boldsymbol{\gamma}_{12}\mathbf{\Gamma}_{11}^{-1}\boldsymbol{I}'_h - \epsilon^2\boldsymbol{I}'_h\mathbf{\Gamma}_{11}^{-11}\boldsymbol{I}_h]$ is positive and then $\det(\tilde{\mathbf{\Gamma}}) > 0$ when $\epsilon \to 0$.

A.4 Theorem 2

For any targeted matrix $\tilde{\mathbf{\Gamma}}$, there exists a positive-definite path from initial $\mathbf{\Gamma}$ that $\mathbf{\Gamma} \xrightarrow{+\eta\mathbf{\Delta}} \tilde{\mathbf{\Gamma}}$, where $\mathbf{\Delta} = \tilde{\mathbf{\Gamma}} - \mathbf{\Gamma}$ and $\eta : 0 \to 1$.

In this positive-definite path, we have

$$\mathbf{\Gamma} + \eta\mathbf{\Delta} = \begin{pmatrix} 1 & \gamma_{12} + \eta\delta_{12} & \cdots & \gamma_{1p} + \eta\delta_{1p} \\ \gamma_{12} + \eta\delta_{12} & 1 & \cdots & \gamma_{2p} + \eta\delta_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{1p} + \eta\delta_{1p} & \gamma_{2p} + \eta\delta_{2p} & \cdots & 1 \end{pmatrix}.$$

Each tiny increment can be decomposed into finite steps implemented by Algorithm 1, i.e.,

$$\begin{aligned} \mathbf{\Gamma} : & \begin{pmatrix} 1 & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & 1 & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & 1 \end{pmatrix} \\ \to & \begin{pmatrix} 1 & \gamma_{12} + \eta\delta_{12} & \cdots & \gamma_{1p} \\ \gamma_{12} + \eta\delta_{12} & 1 & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{1p} & \gamma_{2p} & \cdots & 1 \end{pmatrix} \\ \to \cdots\cdots \to & \begin{pmatrix} 1 & \gamma_{12} + \eta\delta_{12} & \cdots & \gamma_{1p} + \eta\delta_{1p} \\ \gamma_{12} + \eta\delta_{12} & 1 & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{1p} + \eta\delta_{1p} & \gamma_{2p} & \cdots & 1 \end{pmatrix} \\ \to \cdots\cdots \to & \begin{pmatrix} 1 & \gamma_{12} + \eta\delta_{12} & \cdots & \gamma_{1p} + \eta\delta_{1p} \\ \gamma_{12} + \eta\delta_{12} & 1 & \cdots & \gamma_{2p} + \eta\delta_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{1p} + \eta\delta_{1p} & \gamma_{2p} + \eta\delta_{2p} & \cdots & 1 \end{pmatrix} \\ = & \mathbf{\Gamma} + \eta\mathbf{\Delta}. \end{aligned}$$

Let $\eta \to 0$ then $\eta\delta_{ij} \to 0$ for all i, j. According to Lemma 1, every step in this decomposition would remain positive-definite. Thus, any symmetric positive-definite matrix with diagonal elements equal to 1 can be attained by Algorithm 2.

Received 21 April 2020

REFERENCES

[1] Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* 1281–1311. MR1804544

[2] Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98**(4) 807–820. MR2860325

[3] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge. MR2061575

[4] Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**(494) 594–607. MR2847973
[5] Cui, Y., Leng, C., and Sun, D. (2016). Sparse estimation of high-dimensional correlation matrices. *Computational Statistics & Data Analysis* **93** 390–403. MR3406221
[6] Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics* **3**(2) 521. MR2750671
[7] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456) 1348–1360. MR1946581
[8] Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal* **19**(1) C1–C32. MR3501529
[9] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3) 432–441.
[10] Gan, L., Narisetty, N. N., and Liang, F. (2019). Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association* **114**(527) 1218–1231. MR4011774
[11] Jagannathan, R. and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance* **58**(4) 1651–1683.
[12] Khondker, Z. S., Zhu, H., Chu, H., Lin, W., and Ibrahim, J. G. (2013). The Bayesian covariance lasso. *Statistics and Its Interface* **6**(2) 243–259. MR3066689
[13] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37**(6B) 4254. MR2572459
[14] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**(2) 365–411. MR2026339
[15] Levine, R. A., Yu, Z., Hanley, W. G., and Nitao, J. J. (2005). Implementing random scan Gibbs samplers. *Computational Statistics* **20**(1) 177–196. MR2162541
[16] Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance* **7**(1) 77–91. MR0103768
[17] Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science* **26**(3) 369–387. MR2917961
[18] Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika* **99**(3) 733–740. MR2966781
[19] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515. MR2417391
[20] Wang, H. (2014). Coordinate descent algorithm for covariance graphical lasso. *Statistics and Computing* **24**(4) 521–529. MR3223538
[21] Wang, H. et al. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis* **7**(4) 867–886. MR3000017
[22] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**(1) 19–35. MR2367824
[23] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**(2) 894–942. MR2604701

Chaojie Wang
Faculty of Science
Jiangsu University
Zhenjiang
China
E-mail address: wang910930@163.com
Xiaodan Fan
Department of Statistics
The Chinese University of Hong Kong
Shatin, N.T.
Hong Kong
E-mail address: xfan@cuhk.edu.hk