Block-diagonal precision matrix regularization for ultrahigh dimensional data

Yihe Yang[a], Hongsheng Dai[b], Jianxin Pan[c,*]

[a]*Mathematical College, Sichuan University, Chengdu 610065, China*
[b]*Department of Mathematical Sciences, University of Essex*
[c]*Department of Mathematics, University of Manchester, Oxford Road, Manchester M13 9PL, UK*

Abstract

A method that estimates the precision matrix of multiple variables in the extreme scope of "ultrahigh dimension" and "small sample-size" is proposed. Initially, a covariance column-wise screening method is provided in order to identify a small sub-group, which are significantly correlated, from thousands and even millions of variables. Then, a regularization of block-diagonal covariance structure of the thousands or millions of variables is imposed, in which only the covariances of variables in that small sub-group are retained and all others vanish. It is further proven that under some mild conditions the vital sub-group identified by the covariance column-wise screening method is consistent. A major advantage of the proposed method is its efficiency - it produces a reliable precision matrix estimator for thousands of variables within a few of seconds while the existing methods take at least several hours and even so still yield inaccurate estimators. Empirical data studies and numerical simulations show that the proposed precision matrix estimation greatly outperforms existing methods in the sense of taking much less computing time and resulting in much more accurate estimation when dealing with ultrahigh dimensional data.

Keywords: Precision matrix estimation, block-diagonal structure, ultrahigh dimensionality.

1. Introduction

Estimation of the precision matrix $\mathbf{\Theta}$ has attracted an increasing amount of attentions in recent years for many statistical problems. In particular, for Gaussian graphical models, the precision matrix characterizes the conditional dependence among certain random variables, which is vital for us to understand their possible causal structures in the form of a graph (Lauritzen, 1996). A good estimation for $\mathbf{\Theta}$ is essential in many scientific areas such as gene expression studies (Cai et al., 2011), financial portfolio investment studies (Fan et al., 2013), social network exploration (Lauritzen, 1996). Nowadays obtaining reliable estimation for $\mathbf{\Theta}$ becomes extremely challenging in these areas because of the ultrahigh dimensionality of datasets studied. Especially, in some specific fields like genetic analysis, the number of individuals (sample-size n) is much smaller than the number of variables (dimension p), making the estimation of precision matrix extremely challenging.

*Corresponding author
Email address: jianxin.pan@manchester.ac.uk (Jianxin Pan)

The sparsity constraint is one of the main ideas used in the literature to estimate $\mathbf{\Theta}$ under high dimensionality settings. The ℓ_1-penalized optimization, originally pioneered by Tibshirani (1996), is a standard tool to impose sparsity on the precision matrix estimator. Yuan and Lin (2007) proposed the penalized maximum likelihood estimation (MLE) of $\mathbf{\Theta}$ and Rothman et al. (2008) and Ravikumar et al. (2011) further analysed the convergence rate and sign consistency of the penalized MLE, respectively. Zhang and Zou (2014) suggested estimating $\mathbf{\Theta}$ through the so-called penalized D-trace loss function. Using the connection between partial correlation coefficients and regression coefficients in Gaussian graph model, Meinshausen and Bühlmann (2006) proposed a neighborhood selection approach to estimate $\mathbf{\Theta}$ through column by column. Yuan (2010) turned the lasso regularizer (Tibshirani, 1996) in the neighborhood selection to Dantzig selector (Candes and Tao, 2007). Cai et al. (2011) provided an alternative column-wise estimation of $\mathbf{\Theta}$, which is known as constrained ℓ_1-minimization for inverse matrix estimation (CLIME). Liu and Luo (2015) further improved the CLIME for better computational performance.

Another major breakthrough was made by Friedman et al. (2008), which combined the penalized MLE and the neighborhood selection approach and introduced a coordinate descent algorithm known as graphical lasso (glasso) in order to find the MLE of $\mathbf{\Theta}$ efficiently. Fan et al. (2009) demonstrated via statistical theories and numerical examples that the bias resulted by the graphical lasso algorithm can be attenuated when using concave penalty (Fan and Li, 2001) and adaptive lasso (Zou, 2006) rather than lasso. Moreover, there is a vast of literature that focused on improving the computational performance of glasso. For example, Witten et al. (2011) and Mazumder and Hastie (2012) presented a necessary and sufficient condition that may identify the block diagonal structure of $\mathbf{\Theta}$, which leads to significant computational improvement because it only needs to apply graphical lasso to each block of the block-diagonal matrix.

This paper concentrates on a more extreme scope of precision matrix estimation, for example, a genetic study with data having several thousands and even millions of gene expressions but with a sample-size of just several dozens. Almost all the aforementioned methods are struggling to solve problems in this more extreme scope. For instance, the ℓ_1-penalized estimations such as the glasso and Witten's method need to specify a tuning parameter in lasso penalty, but the empirical rule such as cross-validation for choosing the tuning parameter is unstable when the sample-size is small. Another issue is that there are multiple saddle points in the ℓ_1-penalized optimization when sample-size is small, which makes the associated algorithm numerically unstable. Moreover, the computing time of the existing algorithms increases dramatically as the dimension of $\mathbf{\Theta}$ diverges.

To address these challenges, we propose the block-diagonal (BD) regularization, which identifies a small sub-group of significantly correlated variables from thousands of variables and vanishes all the insignificant correlations of variables outside this sub-group. We then approximate the precision matrix by using a BD matrix composed of two blocks, where the low-dimensional block corresponds to this small sub-group of the significantly correlated variables, while the rest high-dimensional block is diagonal. As a result, the estimation of the high-dimensional precision matrix reduces to the estimation of the low-dimensional block plus the high-dimensional but diagonal block.

An identification procedure called covariance column-wise screening (CCS) is proposed to find the small sub-group of highly correlated variables. Consider a multivariate variable $X = (X_j)_{p\times 1}$ with covariance matrix $\mathbf{\Sigma} =$

$(\Sigma_{ij})_{p\times p}$. The CCS procedure uses the jth screening statistic

$$\varrho_j = \sum_{i\neq j} |\hat{\Sigma}^{\text{ini}}_{ij}|, \tag{1}$$

to indicate the significance of the jth component X_j, where $\hat{\mathbf{\Sigma}}^{\text{ini}} = (\hat{\Sigma}^{\text{ini}}_{ij})_{p\times p}$ is an initial covariance matrix estimator. We retain the dependencies of components that have significantly large screening statistics $\{\varrho_j\}$. This CCS procedure is in spirit of the dependence screening approach, which removes the vast majority of the variables out of the current model and then refines the analysis as the second stage in the extreme scope of ultrahigh dimension and small sample-size. Various correlation statistics, such as Pearson correlation coefficient (Fan and Lv, 2008), distance correlation coefficient (Li et al., 2012), and ball correlation coefficient (Pan et al., 2019), can be employed to measure the importance of certain variables. The statistical interpretation of ϱ_j is clear: the expectation of ϱ_j is the sum of non-diagonal entries in the jth column of $\mathbf{\Sigma}$. If a component X_j is independent of the others, ϱ_j has an expectation of zero. Hence, ϱ_j is an appropriate quantity to measure the importance of the jth component X_j in our BD regularization.

Compared with the existing approaches, the BD regularization has the following advantages. First, the BD regularization is not sensitive to the choice of involved tuning parameters. Whereas, almost all the precision matrix estimations based on ℓ_1-penalization encounter numerical instability, due to the fact that different tuning parameters in the ℓ_1-penalty may lead to different results. Besides, the BD regularization is able to produce a consistent precision matrix estimator, while the covariance-insured screening method (He et al., 2019) only answers which entries in the precision matrix are non-zero. Furthermore, our method can identify the sub-group of highly correlated variables by using an initial covariance matrix estimator $\hat{\mathbf{\Sigma}}^{\text{ini}}$. Although Fan and Kim (2019) proposed an analogous method to our BD regularization, their method has a major constraint that the sub-group of highly correlated variables must be known in advance.

The rest of this paper is organized as follows. In Section 2, we introduce the BD regularization of the precision matrix in detail and investigate the CCS identification procedure. In Section 3, we provide the simulation studies. In Section 4, we apply the proposed approach to a real data analysis. Discussion is given in section 5, and all technical proofs are relegated to the Appendix.

2. Methodology

Throughout the paper we use the following notations. Consider a multivariate random variable $X = (X_1, \ldots, X_p)^\top$ that has mean $\mu = (\mu_1, \ldots, \mu_p)^\top$ and covariance matrix $\mathbf{\Sigma} = (\Sigma_{ij})_{p\times p}$. The precision matrix is defined by $\mathbf{\Theta} = (\Theta_{ij})_{p\times p} = \mathbf{\Sigma}^{-1}$. Besides, assume $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ are n independently identically distributed (i.i.d.) random samples from X and denote the sample covariance matrix by

$$\mathbf{S} = (n-1)^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top, \tag{2}$$

where $\bar{\mathbf{X}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i$. For a vector $a = (a_j)_{p\times 1}$, denote $||a||_1 = \sum_j |a_j|$, $||a||_2^2 = \sum_j a_j^2$, $||a||_\infty = \max_j |a_j|$. For a symmetric matrix $\mathbf{A} = (A_{ij})_{p\times p}$, denote $\mathbf{A}^\top$ to be the matrix transpose operator, $||\mathbf{A}||_1 = \max_i \sum_j |A_{ij}|$, $||\mathbf{A}||_2 = \sigma_1(\mathbf{A})$ and $\|\mathbf{A}\|_* = \sum_j \sigma_j(\mathbf{A})$, where $\sigma_j(\mathbf{A})$ is the jth largest eigenvalue of matrix $\mathbf{A}$. We use $\mathrm{E}(\cdot)$ to denote the expectation operator and $\Pr(\cdot)$ to denote the probability operator. We write $a_n \asymp b_n$ if there are positive constants c and C such that $c \le a_n/b_n \le C$. Write $a_n \lesssim b_n$ if there exists a constant C such that $a_n \le Cb_n$. Operators $O(\cdot)$, $o(\cdot)$ are the infinitely large quantity and infinitely small quantity, and $O_P(\cdot)$, $o_P(\cdot)$ denotes that relationships hold with probability tending to 1. Notations diag($\mathbf{A}$) and vec($\mathbf{A}$) are the diagonalizing operator and vectorizing operator of $\mathbf{A}$, respectively. Furthermore, the scale of a set $\mathcal{M}$, i.e., the number of elements in $\mathcal{M}$, is defined as $\#\mathcal{M}$.

2.1. Block-diagonal Regularization of Precision Matrix

In many ultrahigh dimensional statistical problems, a random variable $X = (X_j)_{p\times 1}$ can be re-sorted as $(X_{\mathcal{M}}, X_{\mathcal{M}^c})^\top$, and especially, only the components in $\mathcal{M}$ are significantly correlated. As a consequence, the precision matrix $\mathbf{\Theta}$ can be rewritten as

$$\mathbf{\Theta} = \begin{pmatrix} \mathbf{\Theta}_{\mathcal{M}\mathcal{M}} & \mathbf{\Theta}_{\mathcal{M}\mathcal{M}^c} \\ \mathbf{\Theta}_{\mathcal{M}^c\mathcal{M}} & \mathbf{\Theta}_{\mathcal{M}^c\mathcal{M}^c} \end{pmatrix}, \tag{3}$$

where the absolute values of the off-diagonal entries in $\mathbf{\Theta}_{\mathcal{M}\mathcal{M}}$ are substantially larger than the ones in the rest sub-matrices. The essence of our method is to approximate the above precision matrix through the following BD structured matrix

$$\mathcal{B}(\mathbf{\Theta}) = \begin{pmatrix} \mathbf{\Theta}_{\mathcal{M}\mathcal{M}} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(\mathbf{\Theta}_{\mathcal{M}^c\mathcal{M}^c}) \end{pmatrix}. \tag{4}$$

We call (4) the BD regularization of (3). In general, the size of $\mathcal{M}$ is much smaller than p, which allows to yield a reliable estimator $\hat{\mathbf{\Theta}}_{\mathcal{M}\mathcal{M}}$ by using the more refined methods such as the glasso and CLIME. We then propose to approximate the precision matrix estimator by

$$\mathcal{B}(\hat{\mathbf{\Theta}}) = \begin{pmatrix} \hat{\mathbf{\Theta}}_{\mathcal{M}\mathcal{M}} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(\mathbf{S}_{\mathcal{M}^c\mathcal{M}^c})^{-1} \end{pmatrix}, \tag{5}$$

for real applications with extremely high dimension and small sample-size.

The BD regularization is similar to the diagonal-covariance (DC) regularization (Bickel and Levina, 2004), both of which originate from the fact that it is almost impossible to explore the precise dependence of ultrahigh dimensional variable X when the sample-size is very small. However, by borrowing information from low-dimensional but dominant sub-components $X_{\mathcal{M}}$, the BD regularization makes it possible. In other words, if the low-dimensional sub-components $X_{\mathcal{M}}$ are significantly correlated while the high-dimensional sub-components $X_{\mathcal{M}^c}$ are almost independent, exploring the inner correlations of $X_{\mathcal{M}}$ obtains most correlation information of X. From a practical point

Algorithm 1 Covariance Column-Wise Screening Procedure

Require:

A 1-norm consistent covariance matrix estimator $\hat{\mathbf{\Sigma}}^{\text{ini}} = (\hat{\Sigma}_{ij}^{\text{ini}})_{p\times p}$;

A threshold ζ.

Ensure:

Calculate screening statistics $\varrho_j = \sum_{i\neq j} |\hat{\Sigma}_{ij}^{\text{ini}}|$ column by column;

Employ ζ to divide $\varrho = (\varrho_1, \ldots, \varrho_p)^\top$ into two non-intersection subvectors $\varrho_{\hat{\mathcal{M}}}$ and $\varrho_{\hat{\mathcal{M}}^c}$ with $\hat{\mathcal{M}} = \{j,\ \varrho_j > \zeta\}$.

of view, the estimation of $\mathbf{\Theta}$ is equivalent to the estimation of $\mathbf{\Theta}_{\mathcal{M}\mathcal{M}}$. Therefore, the BD regularization enjoys very light computing consumption and concise implementation.

We then need to address the following two key issues, in order to make the BD regularization work well. First, we should provide a procedure to identify the significantly correlated sub-group $\mathcal{M}$ in practice. Second, we need to give theoretical supports that this identification procedure is able to find $\mathcal{M}$ consistently. We address these two issues in the following two subsections, respectively.

2.2. Covariance Column-Wise Screening

The identification procedure is exhibited in Algorithm 1, which we call the covariance column-wise screening (CCS) procedure and it is actually motivated by the sure independence screening (SIS) method (Fan and Lv, 2008). In fact, the SIS stresses that variables are probably not associated with the response in a regression model if their covariances are insignificant. Likewise, the BD regularization emphasizes that the jth component X_j can be regarded as approximately independent of other components if the corresponding screening statistics ϱ_j is relatively small. The essence of the CCS procedure is to find the gap between $\{\varrho_j, j \in \mathcal{M}\}$ and $\{\varrho_j, j \in \mathcal{M}^c\}$. Theoretically, there should exist such a threshold ζ that concretely separates $\min_{j\in\mathcal{M}} \varrho_j$ and $\max_{j\in\mathcal{M}^c} \varrho_j$, if $\mathbf{\Sigma}$ can really be approximated by a block-diagonal matrix.

A 1-norm consistent covariance matrix estimator $\hat{\mathbf{\Sigma}}^{\text{ini}}$ is necessary in the identification procedure, as otherwise the threshold ζ cannot be found theoretically. Here we suggest to take the adaptive thresholding approach (Cai and Liu, 2011) as the first choice for $\hat{\mathbf{\Sigma}}^{\text{ini}}$, because this approach has a cheap computational cost and exhibits high accuracy in practice. The main idea of adaptive thresholding approach is to shrink the insignificant entries of the sample covariance matrix $\mathbf{S}$ through a generalized thresholding operator $T_\tau(\cdot)$, where common choices of $T_\tau(\cdot)$ include the soft-thresholding operator (Tibshirani, 1996) and the smoothly clipped absolute deviation (SCAD, Fan and Li (2001)). The resulting thresholding covariance estimator has a form $\hat{\mathbf{\Sigma}}^{\mathcal{T}} = (\hat{\Sigma}_{ij}^{\mathcal{T}})_{p\times p}$ where

$$\hat{\Sigma}_{ij}^{\mathcal{T}} = T_{\tau_{ij}}(S_{ij})\mathbf{1}(i \neq j) + S_{ii}\mathbf{1}(i = j), \tag{6}$$

and $\mathbf{1}(\cdot)$ is the indicator function. Regarding the penalizing parameter τ_{ij} corresponding to entry S_{ij}, Cai and Liu

(2011) suggested using the adaptive penalizing parameter

$$\tau_{ij}^{\text{ada}} = \gamma\sqrt{\hat{\theta}_{ij} n^{-1}\log p}, \tag{7}$$

where $\hat{\theta}_{ij} = \widehat{\text{var}}((X_i - \mu_i)(X_j - \mu_j))$, which can be estimated by the moment method. Cai and Liu (2011) also provided an alternative penalizing parameter

$$\tau_{ij}^{\text{alt}} = \gamma\sqrt{\hat{\Sigma}_{ii}\hat{\Sigma}_{jj} n^{-1}\log p}, \tag{8}$$

where the empirical estimators $\hat{\Sigma}_{ii}$ and $\hat{\Sigma}_{jj}$ are S_{ii} and S_{jj}, respectively. Compared to the adaptive penalizing parameter, this alternative shares the same accuracy but is easier to obtain. Regarding the choice of γ, one may simply take $\gamma = 2$ as its justification was verified by many empirical studies, see, e.g., Cai and Liu (2011). Fan et al. (2013) proposed an alternative criterion for choosing γ, which finds the minimum of γ that guarantees $\hat{\Sigma}^{\mathcal{T}}$ to be positive definite. Let $\hat{\mathbf{\Sigma}}^{\mathcal{T}}(\gamma)$ be the adaptive thresholding covariance estimator using γ. Then the optimal γ^{opt} is chosen by

$$\gamma^{\text{opt}} = \inf\left\{\gamma,\ \sigma_{\min}(\hat{\mathbf{\Sigma}}^{\mathcal{T}}(\gamma)) > 0\right\}. \tag{9}$$

In general, γ^{opt} is very likely less than 2 when n is large, leading to a covariance matrix estimator with lower shrinkage. On the other hand, γ^{opt} may be greater than 2 when n is small, and in this case a positive definite covariance matrix estimator can be guaranteed. We utilize this criterion in the numerical studies presented later due to its better performance than setting $\gamma = 2$.

2.3. Asymptotic Properties

In this subsection, we prove that the CCS procedure can identify the "true" sub-group of highly correlated variables with probability tending to 1. The technical assumptions are listed below, and the rigorous proofs are provided in the Supplementary Materials.

Assumption 1 (Exponential-type Tails) Suppose $X_i = (X_{i1}, \ldots, X_{ip})^\top$ and $Y_{ij} = (X_{ij} - \mu_j)/\text{var}(X_{ij})^{1/2}$ where $\mu_j = \text{E}(X_{ij})$. There exist positive constants η_0, τ_0 and r_0 such that for $\forall|t| < \eta_0$, $\sup_j \text{E}\left(\exp(tY_{ij}^2)\right) \leq r_0 < \infty$, and $\min_{jk} \text{var}(Y_{ij}Y_{ik}) \geq \tau_0$.

Assumption 2 (Bounded Eigenvalues) Suppose $\Sigma_{jk} = \text{E}((X_{ij} - \mu_j)(X_{ik} - \mu_k))$ and $\mathbf{\Sigma} = (\Sigma_{jk})_{p\times p}$. There exists a constant ε_0 independent of p such that $0 < \varepsilon_0 \leq \sigma_p(\mathbf{\Sigma}) \leq \sigma_1(\mathbf{\Sigma}) \leq \varepsilon_0^{-1} < +\infty$.

Assumption 1 relaxes the Gaussian distribution restriction to a moment inequality. In the literature, the distribution satisfying this assumption is called a sub-Gaussian distribution (Cai and Liu, 2011). Assumption 2 describes that the covariance matrix $\mathbf{\Sigma}$ is always positive definite no matter how large its dimension is, which plays a central role in large covariance matrix regularization (Bickel and Levina, 2008b).

Assumption 3 (ℓ_q Sparse Covariance) Denote

$$\bar{\delta}_{\mathcal{M}} = \max_{j\in\mathcal{M}} \sum_{i\in\mathcal{M}, i\neq j} |\Sigma_{ij}|^q,\ \bar{\delta}_{\mathcal{M}\mathcal{M}^c} = \max_{j\in\mathcal{M}} \sum_{i\in\mathcal{M}^c} |\Sigma_{ij}|^q,\ \bar{\delta}_{\mathcal{M}^c} = \max_{j\in\mathcal{M}^c} \sum_{i\in\mathcal{M}^c, i\neq j} |\Sigma_{ij}|^q,\ \underline{\delta}_{\mathcal{M}} = \min_{j\in\mathcal{M}} \sum_{i\in\mathcal{M}, i\neq j} |\Sigma_{ij}|^q, \tag{10}$$

with $q \in [0, 1)$. There exist constants C_0 and C_1 larger than 0 such that $\bar{\delta}_{\mathcal{M}} + \bar{\delta}_{\mathcal{M}\mathcal{M}^c} + \bar{\delta}_{\mathcal{M}^c} \leq C_1 < \infty$ and $\max_i \Sigma_{ii} \leq C_0$.

Assumption 4 (ℓ_q Separable Covariance) It is assumed that $\max\{\bar{\delta}_{\mathcal{M}\mathcal{M}^c}, \bar{\delta}_{\mathcal{M}^c}\} = \mathrm{o}(\underline{\delta}_{\mathcal{M}})$, where $\bar{\delta}_{\mathcal{M}\mathcal{M}^c}$, $\bar{\delta}_{\mathcal{M}^c}$ and $\underline{\delta}_{\mathcal{M}}$ are defined in (10).

Assumption 3 indicates that $\mathbf{\Sigma}$ belongs to the thresholdable class of covariance matrix (Bickel and Levina, 2008a). With this assumption, we are able to produce a 1-norm consistent estimator of $\mathbf{\Sigma}$ through certain thresholding approaches (Bickel and Levina, 2008a; Cai and Liu, 2011; Rothman et al., 2009), which is necessary for the identification procedure. Assumption 4 specifies a theoretical gap between $\mathcal{M}$ and $\mathcal{M}^c$, i.e. the minimum ℓ_q column sum of entries in $\Sigma_{\mathcal{M}\mathcal{M}}$ dominates the maximum ℓ_q column sum of entries in $\mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}}$ and $\mathbf{\Sigma}_{\mathcal{M}^c\mathcal{M}^c}$.

Theorem 1 (Consistency of Inputting Covariance Matrix) Let Assumption 1-4 and Condition 1 in the Supplementary Materials hold. If $\log p = o(n^{1/3})$, then

$$\big|\big|\hat{\mathbf{\Sigma}}^{\mathrm{ini}} - \mathbf{\Sigma}\big|\big|_1 = O_P\big((n^{-1}\log p)^{(1-q)/2}\big),$$

where $\hat{\mathbf{\Sigma}}^{\mathrm{ini}}$ is set as the adaptive thresholding covariance estimator $\hat{\mathbf{\Sigma}}^{\mathcal{T}}$ (6).

Theorem 1 indicates that there exists an 1-norm consistent covariance estimator yielded by the adaptive thresholding approach, based on which we can identify the ℓ_q significantly correlated sub-group $\mathcal{M}$. Especially, both the adaptive penalizing parameter $\tau_{ij} = \gamma\sqrt{((n^{-1}\log p)\hat{\theta}_{ij})}$ and alternative penalizing parameter $\tau_{ij} = \gamma\sqrt{((n^{-1}\log p)\hat{\Sigma}_{ii}\hat{\Sigma}_{jj})}$ ensure Theorem 1 must hold.

Theorem 2 (Identifiability of ℓ_q Separable Covariance Structure) Let Assumption 1-4 and Condition 1 in the Supplementary Materials hold, and $(n^{-1}\log p)^{1-q)/2} = o(\underline{\delta}_{\mathcal{M}})$. Then there exists a threshold $\zeta \asymp \underline{\delta}_{\mathcal{M}}$ such that

$$\Pr(\hat{\mathcal{M}} \subset \mathcal{M}) \geq 1 - O((\log p)^{-\frac{1}{2}} p^{-\gamma+2}), \quad \Pr(\mathcal{M} \subset \hat{\mathcal{M}}) \geq 1 - O((\log p)^{-\frac{1}{2}} p^{-\gamma+2}),$$

where $\varrho_j = \sum_{i\neq j} |\hat{\Sigma}^{\mathcal{T}}_{ij}|$, $\hat{\mathcal{M}} = \{j, \varrho_j > \zeta\}$, and γ is the multiplier in penalizing parameters (7) and (8).

Theorem 2 implies that there exists a threshold ζ such that $\Pr(\hat{\mathcal{M}} = \mathcal{M}) \to 1$ as $p \to \infty$. Likewise, both the adaptive one and alternative one are able to guarantee the identification of $\mathcal{M}$. Note that it may be difficult to find ζ analytically, due to the failure of the dependence screening approach. We give two criteria to select ψ, which is an equivalent quantity of ζ, in section 2.4.

Theorem 3 (Consistency of the BD precision matrix estimator) Suppose that Assumption 1-4 are satisfied. If $\max\{\bar{\delta}(\mathcal{M}\mathcal{M}^c), \bar{\delta}(\mathcal{M}^c)\} = o(1)$, then

$$||\mathcal{B}(\hat{\mathbf{\Theta}}) - \mathbf{\Theta}||_2 = O(\delta_\Theta + \bar{\delta}^{\frac{1}{q}}_{\mathcal{M}\mathcal{M}^c} + \bar{\delta}^{\frac{1}{q}}_{\mathcal{M}^c}),$$

where $\delta_\Theta = ||\hat{\mathbf{\Theta}}_{\mathcal{M}\mathcal{M}} - \mathbf{\Theta}_{\mathcal{M}\mathcal{M}}||_2$.

Theorem 3 points out that the estimation error of the BD precision estimator $\mathcal{B}(\hat{\mathbf{\Theta}})$ is $||\mathcal{B}(\hat{\mathbf{\Theta}}) - \mathbf{\Theta}||_2 = O(\delta_\Theta + \bar{\delta}^{1/q}_{\mathcal{M}\mathcal{M}^c} + \bar{\delta}^{1/q}_{\mathcal{M}^c})$, which declines to zero if the correlations outside the sub-group $\mathcal{M}$ are ignorable. Hence, as long as we correctly recover $\mathcal{M}$ through the CCS procedure and apply a well-defined method to yield $\hat{\mathbf{\Theta}}_{\mathcal{M}\mathcal{M}}$, the resulting BD precision estimator $\mathcal{B}(\hat{\mathbf{\Theta}})$ is guaranteed to be consistent. See Ravikumar et al. (2011) and Rothman et al. (2008) for theoretical results of sparse precision matrix estimation.

2.4. Selection Criteria of ψ

We introduce a second method to find $\mathcal{M}$, solving the problem that the cutoff ζ may be difficult to specify in implementation. Sort $\varrho_1, \varrho_2, \ldots, \varrho_p$ as $\varrho_{i_1} \leq \varrho_{i_2} \leq \cdots \leq \varrho_{i_p}$ and select an appropriate integer $\psi \in (1, p)$ such that

$$\hat{\mathcal{M}} = \{i_{p-\psi+1}, i_{p-\psi+2}, \ldots, i_p\}. \tag{11}$$

Given an appropriate ψ, the cutoff ζ can be chosen as any quantity between $\varrho_{i_{p-\psi}}$ and $\varrho_{i_{p-\psi+1}}$. In practice, we need to properly select the number of retained components ψ. A natural procedure is to use cross-validation criterion, however, it is extremely time-consuming in ultrahigh dimensional setting. To avoid this problem, we provide two alternative criteria.

One criterion is to simply set the number of retained components ψ to be $[dn]$ with $d \in (0, 1)$. Indeed, in the SIS and its related extensions, the number of retained variables is usually set to be $[4n/\log p]$. In spirit of the SIS, we suggest to use $[4n/\log p]$ or any other appropriate value to be a selection criterion.

The second criterion is provided as follows. Consider the series of the increasingly ordered screen statistics $\{\varrho_{i_j}\}$ and their differences $\{\mathrm{d}\varrho_{i_1}, \mathrm{d}\varrho_{i_2}, \ldots, \mathrm{d}\varrho_{i_{p-1}}\}$, where $\mathrm{d}\varrho_{i_j} = \varrho_{i_j+1} - \varrho_{i_j}$. We treat $\{\mathrm{d}\varrho_{i_i}\}$ as a random process and then employ a change-point recognizer to find its change-point. The cumulative sum control chart (CUSUM, Page (1954)) is the most common change-point recognizer, of which the principle is to track the cumulative sum $c_0 = 0$ and $c_{i+1} = \max(0, c_i + Z_i - w_i)$, with samples from a general random process $\{Z_1, Z_2, \ldots\}$ and the corresponding pre-assigned weights $\{w_1, w_2, \ldots\}$. When the value of the j^*th cumulative sum c_{j^*} exceeds a certain threshold, the change-point location is detected. The number of retained components is then $\psi = p - j^*$. In practice, we suggest implementing CUSUM on the last s sub-components of $\{\mathrm{d}\varrho_{i_j}\}$ where s may be taken as 100, 200 or n for the simplicity of computation. The choice of s depends on certain prior knowledge, for example, the maximum number of variables that are expected to be included in $\mathcal{M}$. In the most circumstances, it is believed that $s = n$ is sufficient to ensure $\#\mathcal{M} \leq s$.

3. Simulation Studies

3.1. Settings of Simulation

The simulation studies in this section aim to carry out the following analyses: (a) to compare the BD regularization with its competitors in high dimensional setting, and (b) to explore the performance of the BD regularization in ultrahigh dimensional setting. For the former one, we consider $n = 400, 800$ and $p = 200s_p$ where $s_p = (1, 2, 3, 4, 5)$. For the latter study, we take $n = 500, 1000$ and $p = 1000s_p$.

The ℓ_q separable correlation structures are set as follows. Consider the decomposition $\mathbf{\Sigma} = \text{diag}(\mathbf{\Sigma})^{1/2}\mathbf{R}\text{diag}(\mathbf{\Sigma})^{1/2}$ where $\text{diag}(\mathbf{\Sigma})$ is the diagonal variance matrix and $\mathbf{R}$ is the correlation matrix. As for $\text{diag}(\mathbf{\Sigma})$, we set $\Sigma_{ii} = 2\cos(2\pi i/p) + 3$. Also, we consider two specific correlation structures for $\mathbf{R}$. Let $\mathcal{M} = \bigcup_{k=1}^{3}\mathcal{N}_k$ where $\mathcal{N}_1 = \{1, \ldots, 30\}$, $\mathcal{N}_2 = \{31, \ldots, 100\}$ and $\mathcal{N}_3 = \{101, \ldots, 200\}$. Let $\mathcal{M}_1 = \mathcal{N}_1$, $\mathcal{M}_2 = \mathcal{N}_1 \cup \mathcal{N}_2$, and $\mathcal{M}_3 = \mathcal{M}$. Outside $\mathcal{M}$ the correlation coefficients are set to be exactly zero. For structure I we set

$$\mathbf{R}_{\mathcal{M}\mathcal{M}} = \begin{pmatrix} \mathbf{R}_{\mathcal{N}_1\mathcal{N}_1} & \mathbf{R}_{\mathcal{N}_1\mathcal{N}_2} & \mathbf{R}_{\mathcal{N}_1\mathcal{N}_3} \\ \mathbf{R}_{\mathcal{N}_2\mathcal{N}_1} & \mathbf{R}_{\mathcal{N}_2\mathcal{N}_2} & \mathbf{R}_{\mathcal{N}_2\mathcal{N}_3} \\ \mathbf{R}_{\mathcal{N}_3\mathcal{N}_1} & \mathbf{R}_{\mathcal{N}_3\mathcal{N}_2} & \mathbf{R}_{\mathcal{N}_3\mathcal{N}_3} \end{pmatrix} = \begin{pmatrix} \text{cs}(0.7) & 0.3\mathbf{1}\mathbf{1}^\top & 0.1\mathbf{1}\mathbf{1}^\top \\ 0.3\mathbf{1}\mathbf{1}^\top & \text{cs}(0.3) & 0.1\mathbf{1}\mathbf{1}^\top \\ 0.1\mathbf{1}\mathbf{1}^\top & 0.1\mathbf{1}\mathbf{1}^\top & \text{cs}(0.1) \end{pmatrix},$$

and for structure II it is

$$\mathbf{R}_{\mathcal{M}\mathcal{M}} = \begin{pmatrix} \mathbf{R}_{\mathcal{N}_1\mathcal{N}_1} & \mathbf{R}_{\mathcal{N}_1\mathcal{N}_2} & \mathbf{R}_{\mathcal{N}_1\mathcal{N}_3} \\ \mathbf{R}_{\mathcal{N}_2\mathcal{N}_1} & \mathbf{R}_{\mathcal{N}_2\mathcal{N}_2} & \mathbf{R}_{\mathcal{N}_2\mathcal{N}_3} \\ \mathbf{R}_{\mathcal{N}_3\mathcal{N}_1} & \mathbf{R}_{\mathcal{N}_3\mathcal{N}_2} & \mathbf{R}_{\mathcal{N}_3\mathcal{N}_3} \end{pmatrix} = \begin{pmatrix} \text{ma}_2(-0.6, 0.3) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \text{ma}_2(-0.4, 0.2) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \text{ma}_2(-0.2, 0.1) \end{pmatrix},$$

where $\text{cs}(\rho)$ denotes the correlation matrix that is of compound symmetry structure with correlation coefficients ρ, and $\text{ma}_2(\rho_1, \rho_2)$ denotes the correlation matrix that is of order-2 moving-average structure with 1st and 2nd autocorrelation coefficients ρ_1 and ρ_2.

Furthermore, we use entropy loss of two matrices to assess the estimation error, whose representation is

$$L(\mathbf{\Theta}, \hat{\mathbf{\Theta}}) = \text{trace}(\hat{\mathbf{\Theta}}\mathbf{\Theta}^{-1}) - \log\det\hat{\mathbf{\Theta}}\mathbf{\Theta}^{-1} - p,$$

which aims to assess the discrepancy of eigenvalues of two matrices. In addition to the entropy loss, the computing time is also counted when comparing the computational efficiency of the precision matrix estimators. The CUP is Inter(R) Xeon(R) Platinum 8160 2.10GHz with 256 GB memory. The replication runs are 300.

3.2. Investigation of BD regularization with different tuning parameters

We first investigate whether or not the BD regularization is sensitive to the choice of the tuning parameters, i.e., the penalizing parameter in the adaptive thresholding approach and the designated size of the sub-group $\mathcal{M}$. For this purpose, we propose the following model set-up for numerical simulations.

Firstly, We use the adaptive thresholding approach to estimate the initial covariance estimator $\hat{\mathbf{\Sigma}}^{\text{ini}}$, where the SCAD operator, whose expression is given in the Supplementary Materials, is employed to shrink the elements. We

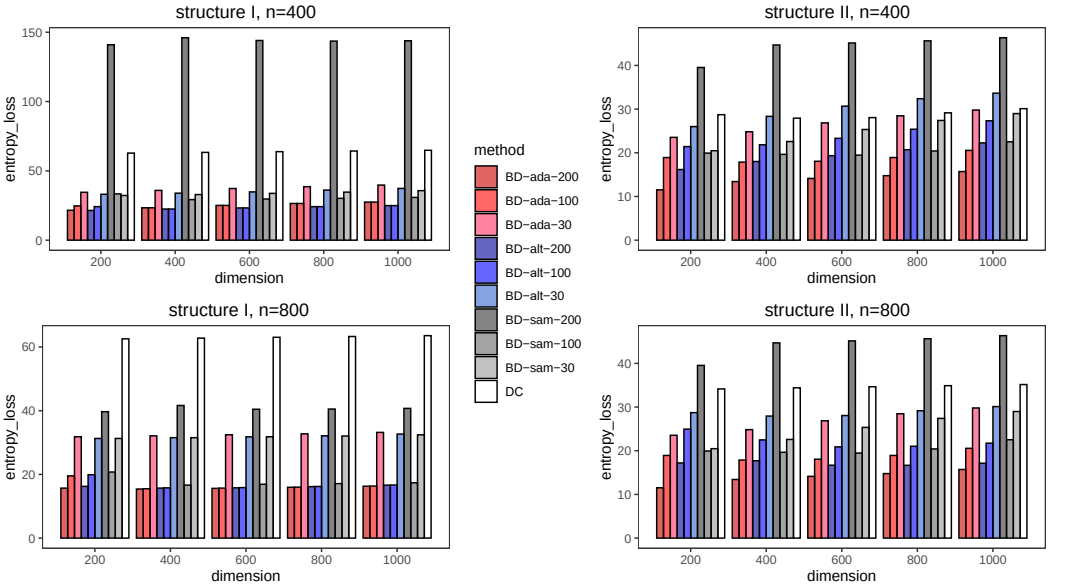

Figure 1: The entropy loss of different BD precision matrix estimators and inverse of diagonal sample covariance matrix. The results for structure I are shown in left two panels and those for structure II are in right two panels. Dimension of precision matrix is $200s_p$ where $s_p = (1, 2, 3, 4, 5)$.

then implement the CCS procedure to find the sub-groups $\hat{\mathcal{M}}_1$, $\hat{\mathcal{M}}_2$ and $\hat{\mathcal{M}}_3$, which respectively consists of the indices of the first 30,100 and 200 largest covariance screen statistics $\{\varrho_j\}$. Without loss of generality, we directly apply the associated BD structure to the initial covariance estimator $\hat{\mathbf{\Sigma}}^{\text{ini}}$ and calculate its inverse as the BD precision matrix estimator. The involved BD precision matrix estimators are as follows

$$\text{BD-ada-30}: \ \mathcal{B}_{\hat{\mathcal{M}}_1}(\hat{\mathbf{\Theta}}^{\mathcal{T}}_{\text{ada}}), \quad \text{BD-ada-100}: \ \mathcal{B}_{\hat{\mathcal{M}}_2}(\hat{\mathbf{\Theta}}^{\mathcal{T}}_{\text{ada}}), \quad \text{BD-ada-200}: \ \mathcal{B}_{\hat{\mathcal{M}}_3}(\hat{\mathbf{\Theta}}^{\mathcal{T}}_{\text{ada}}),$$
$$\text{BD-alt-30}: \ \mathcal{B}_{\hat{\mathcal{M}}_1}(\hat{\mathbf{\Theta}}^{\mathcal{T}}_{\text{alt}}), \quad \text{BD-app-100}: \ \mathcal{B}_{\hat{\mathcal{M}}_2}(\hat{\mathbf{\Theta}}^{\mathcal{T}}_{\text{alt}}), \quad \text{BD-alt-200}: \ \mathcal{B}_{\hat{\mathcal{M}}_3}(\hat{\mathbf{\Theta}}^{\mathcal{T}}_{\text{alt}}),$$
$$\text{BD-sam-30}: \ \mathcal{B}_{\hat{\mathcal{M}}_1}(\mathbf{S}), \quad \text{BD-sam-100}: \ \mathcal{B}_{\hat{\mathcal{M}}_2}(\mathbf{S}), \quad \text{BD-sam-200}: \ \mathcal{B}_{\hat{\mathcal{M}}_3}(\mathbf{S}),$$

where $\mathcal{B}_{\hat{\mathcal{M}}}(\hat{\mathbf{\Theta}}^{\mathcal{T}}_{\text{ada}}) = \{\mathcal{B}_{\hat{\mathcal{M}}}(\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{ada}})\}^{-1}$, with $\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{ada}}$ estimated from the thresholding approach with adaptive penalizing parameters (7); $\mathcal{B}_{\hat{\mathcal{M}}}(\hat{\mathbf{\Theta}}^{\mathcal{T}}_{\text{alt}}) = \{\mathcal{B}_{\hat{\mathcal{M}}}(\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{alt}})\}^{-1}$, with $\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{\text{alt}}$ estimated from the thresholding approach with alternative penalizing parameters (8). We use the BD-sam, i.e., the BD regularization with sample covariance matrix input in CCS, to confirm that it is necessary to use the thresholding covariance estimators rather than the sample covariance matrix. In addition, the competitors of the BD regularization are DC regularization.

The performance of the BD regularization when inputting different covariance matrix estimators in CCS are studied subsequently. Figure 1 shows the dynamic changes of entropy loss of the different BD precision matrix estimators as the dimension increases from 200 to 1000. From this figure, we give two comments below.

We first investigate whether or not the choice of penalizing parameter τ_{ij} corresponding to entry S_{ij} has effects on the BD regularization. For structure I, the adaptive scheme (7) and alternative scheme (8) share little difference. However, for structure II, the former performs uniformly better than the latter. Cai and Liu (2011) has commented that

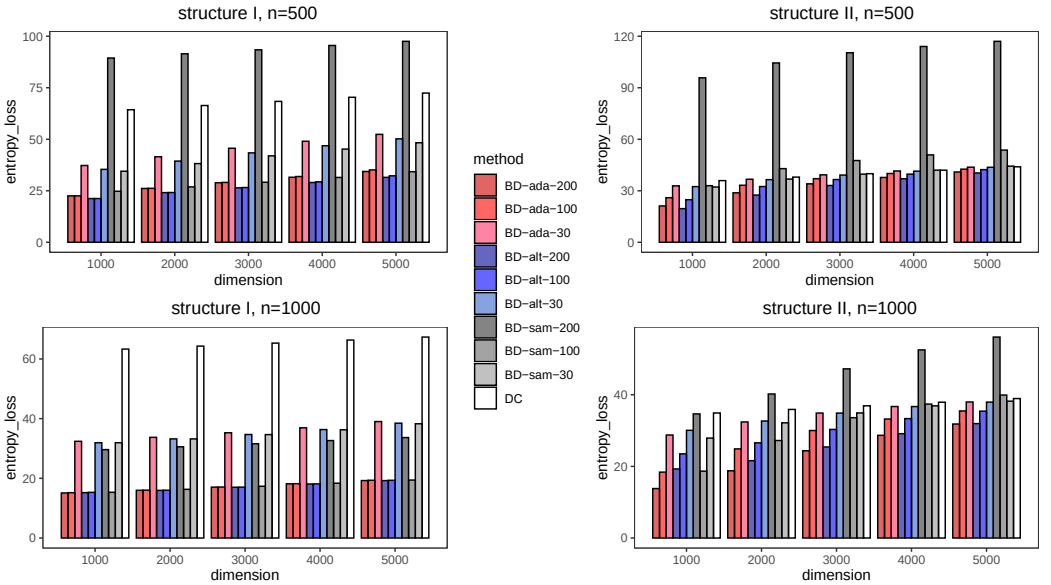

Figure 2: The entropy loss of different BD precision matrix estimators and inverse of diagonal sample covariance matrix. The results for structure I are shown in left two panels and those for structure II are in right two panels. Dimension of precision matrix is $1000s_p$ where $s_p = (1, 2, 3, 4, 5)$.

the adaptive scheme is better than the alternative scheme, because it is difficult to choose γ in the alternative scheme, i.e. $\tau_{ij}^{\text{alt}} = \gamma\sqrt{(\hat{\Sigma}_{ii}\hat{\Sigma}_{jj}n^{-1}\log p)}$.

The BD-sam may provide a good precision matrix estimator, provided that the cardinality of $\mathcal{M}$ is chosen reasonably. However, when the cardinality is inappropriately chosen, e.g., $\#\hat{\mathcal{M}} = 200$, the BD-sam method may identify too many false indices in $\mathcal{M}$ and consequently performs worse than the DC regularization. This situation is somewhat alleviated if the sample-size is boosted. Even so, it still remains the worst, and hence inputting a 1-norm consistent covariance matrix is really necessary.

3.3. Investigation of BD regularization in ultra-high dimension

We now turn to study if the BD regularization still performs well under the ultra-high-dimensional setting. Figure 2 shows the dynamic changes of entropy loss of different BD precision matrix estimates when the dimension increases from 1000 to 5000. We slightly boost the sample-size to $n = 500, 1000$ because the dimension is now very high.

It is clear from Figure 2 that, for ultrahigh dimensional precision matrix estimation, the BD regularization is reliable and performs significantly better than the DC regularization. It is because the BD regularization retains significant entries in a 1-norm consistent covariance matrix estimate, i.e., it maintains the vital correlations of the ultrahigh dimensional random variable. On the other hand, the exact precision matrix estimating methods, such as the glasso and neighborhood selection, can hardly handle such an ultrahigh dimensional precision matrix. In contrast, the BD regularization still works very well because it enforces certain entries into a small sub-matrix and sets all the off-diagonal entries outside this sub-matrix to be zero. Therefore, calculating the inverse of $\mathbf{\Sigma}$ is equivalent to

calculating the inverse of the small matrix $\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}$, so that it becomes an easy task to do.

The adaptive scheme (7) and alternative scheme (8) have little difference for most cases, although Cai and Liu (2011) proved that the former is better in theory. Actually, in the implementation of the adaptive scheme, we employ the moment method to estimate $\hat{\theta}_{ij}$ in the adaptive penalizing parameter $\tau_{ij}^{\text{ada}} = \gamma\sqrt{(\hat{\theta}_{ij} n^{-1} \log p)}$. When $p \gg n$, this moment estimator inevitably causes a great estimation error and consequently worsens the performance of the adaptive scheme. Moreover, it takes considerable time to calculate the moment estimator $\hat{\theta}_{ij}$, although it may be non-iterative in general. Therefore, we suggest the alternative scheme if computational costs are a primary issue.

3.4. Comparison between BD regularization and the competitors

In this subsection, we compare the BD regularization with commonly used competitors including the DC regularization (Bickel and Levina, 2004), glasso (Friedman et al., 2008), CLIME (Cai et al., 2011), and the adaptive thresholding approach which directly calculates the inverse of the resulting covariance esitmator (Cai and Liu, 2011). The glasso and CLIME are respectively implemented through R packages "glasso" and "fastclime" in R 4.1.2, where the optimal tuning parameters are determined by Bayesian information criterion (BIC, Schwarz (1978)). On the other hand, as mentioned in the previous subsections that the adaptive and alternative schemes differ only slightly in identifying the sub-group $\mathcal{M}$, for convenience we use the alternative scheme to identify $\mathcal{M}$ and use the adaptive scheme to estimate the covariance matrix $\hat{\mathbf{\Sigma}}_{\mathcal{M}\mathcal{M}}$. This strategy improves the BD regularization in terms of computational efficiency, because calculating the moment estimator $\hat{\theta}_{ij} = \widehat{\text{var}}((X_i - \mu_i)(X_j - \mu_j))$ one-by-one becomes costly when p is large. Furthermore, we combine the BD regularization with the glasso, that is, firstly identify $\mathcal{M}$ using the CCS procedure with the alternative scheme, and then obtain $\hat{\mathbf{\Theta}}_{\mathcal{M}\mathcal{M}}$ using the glasso, where the optimal tuning parameter in the glasso is also determined by BIC. Although this glasso-based BD regularization may be more time-consuming, it yields a more accurate precision matrix estimator. For these two BD regularizations, we implement the CUSUM to the first 500 differences $\{\mathrm{d}\rho_{i_j}\}$ in order to determine the optimal cutoff ψ (see details in section 2.4).

The outputs for the comparisons between the BD regularization and the competitors are shown as follows. Figure 3 shows the dynamic changes of entropy loss and computing time of the BD regularization and its competitors when p increases from 550 to 1000 and n set as 400. From this figure, we have observed the following findings.

First, with both precision matrix structures, the glasso-based BD regularization performs the best in terms of estimation error. It even outperforms the theoretically most accurate glasso, demonstrating that it is capable of producing a more accurate precision matrix estimator by incorporating more structural information. Besides, the BD regularization with the adaptive thresholding approach behaves well, and its estimation error is lower than when the adaptive thresholding approach is used directly. Furthermore, in our simulations, the CLIME performs poorly: it is only slightly better than the DC regularization and significantly worse than other methods. This poor performance could be due to a lack of maintenance and updates in the R package "fastclime." We believe that we will achieve a better result if we carefully implement the CLIME ourselves, yet this result is no better than the glasso.

Second, the BD regularizations with the glasso and adaptive thresholding approach are much more efficient than

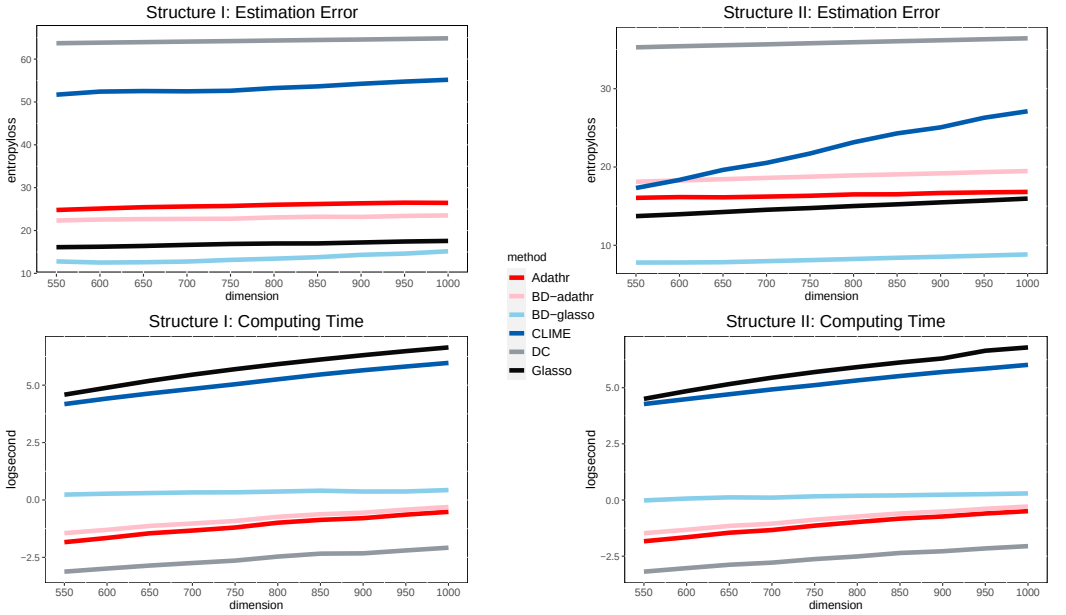

Figure 3: The dynamic changes of entropy loss and computing of the BD regularization and its competitors. The notations "BD-adathr" and "BD-glasso" refer to the BD regularization with the adaptive thresholding approach and glasso, respectively.

the glasso and CLIME in terms of computation. We present the logarithm of the number of seconds because the numbers of seconds taken by these two groups of methods are not in the same order of magnitude. It should be noted that the precision matrix in the simulations has a dimension of no more than 1000, therefore it does not fall within the scope of the ultra-high dimension. When the glasso and CLIME are applied to the ultra-high-dimensional scope where the precision matrix has a dimension of at least several thousand, neither of them can converge in a reasonable amount of time. This numerical evidence illustrates that the BD regularization, particularly the glasso-based BD regularization, has a great deal of potential for use in ultra-high dimensional applications.

3.5. Comparison in misspecificed case

Now we investigate the performance of the BD regularization when the BD structure is misspecified. We generate the misspecified structure by increasing the sizes of the second and third index sets: $\mathcal{N}_2 = \{31+q, \ldots, 100+q\}$ and $\mathcal{N}_3 = \{101+q, \ldots, 200+q\}$ where q starts from 20 and ends at 200. As q increases, the resulting structure of the covariance matrix gradually violates the BD structure – there exists a small sub-group that contains the highly correlated variables. The dimension p is set 800 and the sample size n is set 400.

The outputs for the comparisons between the BD regularization and the competitors are shown as follows. Figure 4 shows the dynamic changes of entropy loss and computing time of the BD regularization and its competitors when q increases from 220 to 400. From this figure, we have observed the following findings.

For structure I, the BD regularization, particularly the glasso-based BD regularization, also suffers from poor performance when q is large. This is because the identified sub-group $\hat{\mathcal{M}}$ differs so much from the true one. In

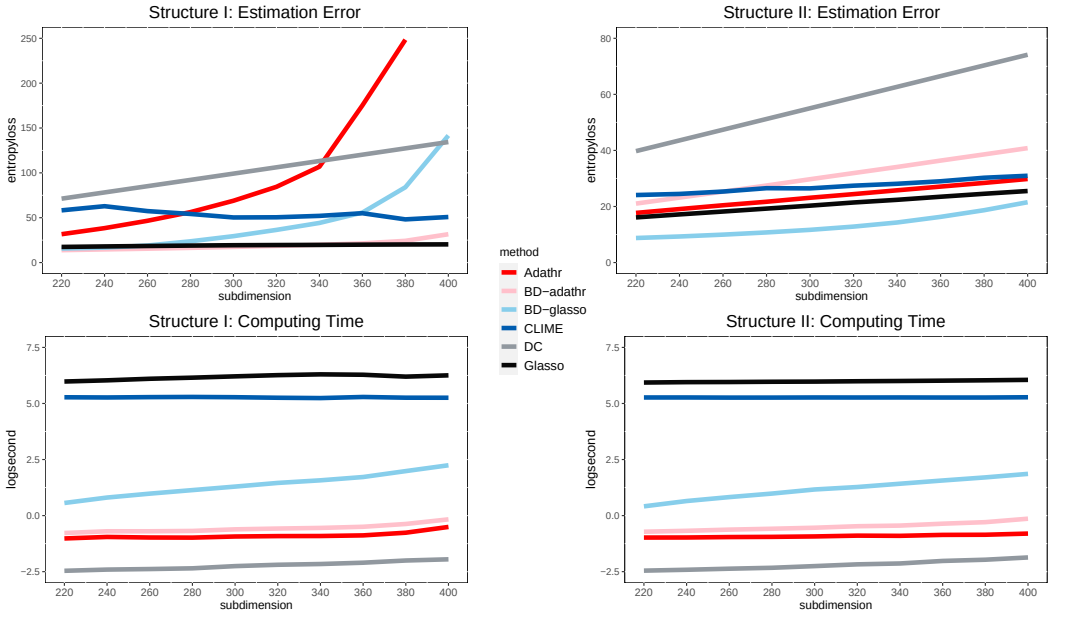

Figure 4: The dynamic changes of entropy loss and computing of the BD regularization and its competitors under misspecified structures. The notations "BD-adathr" and "BD-glasso" refer to the BD regularization with the adaptive thresholding approach and glasso, respectively.

contrast, the glasso and CLIME behave stably and are insensitive to the change of q, since they do not impose any other special structures on the precision matrix besides the sparsity structure. In addition, the adaptive thresholding approach cannot produce a positive definite estimator when p is large and the underlying structure of the covariance matrix is complete. Such a problem further causes the entropy loss between it and the true precision matrix is indeed an imaginary number with a very large real part. As a consequence, it performs dramatically poorly for large q.

As for structure II, the BD regularization outperforms the others, regardless of how large q is. It could be because $\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}$ is the covariance matrix of a stationary process in structure II, ensuring that the leading eigenvalues of $\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}$ change smoothly as q rises. Such a property makes it easier to estimate the corresponding precision matrix than the one under structure I. Hence, the BD regularization is not affected by the violation of the structure II.

In conclusion, the BD regularization is robust to the misspecified structure if this misspecified one is not complex, e.g., the structure II under which $\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}$ is also highly sparse. Whereas, when the misspecified structure is very complex, for example, the size of $\mathcal{M}$ is large and at the same time $\mathbf{\Sigma}_{\mathcal{M}\mathcal{M}}$ is not sparse, the BD regularization is expected to perform poorly. The reason for this breakdown is due to that the identified sub-group $\hat{\mathcal{M}}$ is considerably different from $\mathcal{M}$. In this case, the traditional methods such as the glasso and CLIME are suggested, although they may take substantially more time to complete the estimation of the precision matrix.

4. Real Application

There are many applications that benefit from the BD regularization, such as discriminant analysis (Bickel and Levina, 2004), portfolio allocation (Fan et al., 2013), network recovery (Cai et al., 2011), etc. In this section, we focus on the ultrahigh dimensional linear discriminant analysis (LDA) and discuss other topics in the Supplementary Materials.

4.1. Ultrahigh Dimensional Linear Discriminant Analysis

Assume that there exist K different p-variate distributions with mean $\mu^{(k)}$ and common precision matrix $\mathbf{\Theta}$, and a label-missing sample $X = (X_j)_{p\times 1}$ generated from one of the K different distributions. We treat k^* as the label of X, i.e. this sample is considered from the k^*th distribution, if k^* minimizes the LDA score

$$\delta_k(X) = -X^\top \mathbf{\Theta}\mu^{(k)} + \frac{1}{2}(\mu^{(k)})^\top \mathbf{\Theta}\mu^{(k)} - \log \pi^{(k)}, \tag{12}$$

where $\pi^{(k)}$ is the prior probability of category k. In practice, we may have the training-set with K categories and aim to discriminate the categories of a new individual in the test-set.

In high-dimensional setting, $\mathbf{\Theta}$ is hard to estimate if $n = \sum_{k=1}^K n_k$ is comparatively smaller than p, where n_k is the sample-size of category k. For resolving this issue, Bickel and Levina (2004) proposed the DC regularization that assumes all the features are independent within in each category. Suppose that $\mathbf{S} = (S_{ij})_{p\times p}$ is the common sample covariance matrix for all categories, then the DC-LDA for category k has the score

$$\delta_k^{\text{DC}}(X) = \sum_{j=1}^{p} \left\{ \frac{-2X_j^\top \hat{\mu}_j^{(k)} + (\hat{\mu}_j^{(k)})^\top \hat{\mu}_j^{(k)}}{2S_{jj}} \right\} - \log \pi^{(k)}, \tag{13}$$

where $\hat{\mu}_j^{(k)}$ is the sample mean of jth component for category k. Hastie et al. (2009) provided a detailed introduction about this method when applied to genetic research.

We stress that the discrimination efficiency can be greatly enhanced if we retain the significant entries of the sample covariance matrices because the corresponding components of X are actually correlated. Hence, we employ the BD regularization and propose a new BD-LDA method whose discriminant score for category k is

$$\delta_k(X)^{\text{BD}} = -X^\top \mathcal{B}(\hat{\mathbf{\Theta}})\hat{\mu}^{(k)} + \frac{1}{2}(\hat{\mu}^{(k)})^\top \mathcal{B}(\hat{\mathbf{\Theta}})\hat{\mu}^{(k)} - \log \pi^{(k)}. \tag{14}$$

This discriminant score can be efficiently calculated, since the significantly correlated block $\hat{\mathbf{\Theta}}_{\mathcal{M}\mathcal{M}}$ is of low dimension.

4.2. Real Data Analysis

We now demonstrate that our BD regularization is able to enhance discrimination efficiency through a real data example. The breast cancer data was originally studied by Hess et al. (2006) and is available at the UCSC Xena (`https://xenabrowser.net/datapages/`). It includes 21816 gene expressions of 133 samples, where 34

samples are with pathological complete response (pCR) and 99 samples are with residual disease (RD). Our aim is to predict whether or not a subject belongs to the pCR state by BD-Adathr-LDA, BD-Glasso-LDA, DC-LDA, and the LDA approaches with the precision matrix estimators yielded by the glasso (glasso-LDA) and CLIME (Clime-LDA). These methods are implemented in the same way that they were in simulation.

To fairly compare these methods, we choose the same scheme as the one used by Fan et al. (2009) and Cai et al. (2011). Specifically, the data are randomly divided into a train-set and a test-set, where 5 pCR samples and 16 RD samples constitute the test-set and the rest samples form the train-set. A two-sample t test is then performed between pCR and RD for each gene to pre-drop partial genes from the discrimination model. For the glasso and CLIME, the 113 most significant genes (with the smallest p-values) are retained. As to the BD regularization and DC regularization, the 20×133 most significant genes are retained where 133 is the number of samples of this data. Next, we calculate the sample mean $\hat{\mu}_k$ for category $k = 1, 2$ and estimate the common sample covariance matrix $\mathbf{S}$. The prior probabilities π_1, π_2 are set to be 34/133 and 99/133, respectively. More details of this real data analysis can be found in the Supplementary Materials.

We propose to use the specificity, sensitivity, and Mathews' correlation coefficient (MCC), computing time, and the number of retained genes to evaluate the discrimination powers for the aforementioned discriminators. The specificity, sensitivity, and MCC criteria are defined as:

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad \text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{MCC} = \frac{\text{TN} \cdot \text{TP} - \text{FN} \cdot \text{FP}}{\sqrt{(\text{TN} \cdot \text{TP})(\text{TN} \cdot \text{FP})(\text{FN} \cdot \text{TP})(\text{FP} \cdot \text{TP})}},$$

where TP and TN denote the true positive (i.e., pCR) and true negative (i.e., RD), and FP and FN refer to false positive and false negatives. The larger the three criteria are, the better the discrimination power is. It is worth to mention that the number of samples in pCR is significantly smaller than the one in RD, which makes the decision boundary shift to RD. Thus, the specificity is likely to be larger than the sensitivity for the breast cancer data.

The results of the discriminant analysis on breast cancer data are presented in Table 1. It shows that BD-adathr-LDA, BD-glasso-LDA, and glasso-LDA have the top discrimination performances in terms of specificity criterion. The methods of CLIME-LDA is even worse than the DC-LDA because this approach is based on the approximate estimators of the precision matrix. As for the sensitivity criterion, the proposed BD-glasso-LDA and BD-adathr-LDA become the best discriminators, while the glasso-LDA suffers from a great reduction of discrimination power. The reason why the glasso-LDA and CLIME-LDA perform so poorly is that these methods removed too many important genes in the two-sample t-test. In fact, when the samples are highly unbalanced, some genes with small discrimination effects are likely to be removed by the two-sample t-test. Accordingly, these three methods encounter the worse indices of sensitivity. For the MCC criterion, the proposed BD-glasso-LDA and BD-adathr-LDA still enjoys the top performance. In addition, although the BD-adathr-LDA handles the discrimination model with a relatively large dimension, it takes nearly the same time when estimating the precision matrix as the glasso-LDA and CLIME-LDA. As for BD-glasso-LDA, its computing time is still within the acceptable range. It indicates that our BD regularization can efficiently deal with the precision matrix estimation in the extreme case of ultrahigh dimension.

Table 1: Comparison of the averaged discrimination error over 100 replications (standard error).

Method	Specificity	Sensitivity	MCC	Computing Time	Dimension
BD-adathr-LDA	0.6912(0.1335)	0.3542(0.2186)	0.0442(0.2348)	2.8707sec(0.1057)	**2660**
BD-glasso-LDA	0.6960(0.1352)	**0.3840**(0.2299)	**0.0779**(0.2386)	10.6075sec(0.9621)	**2660**
DC-LDA	0.6764(0.1271)	0.3326(0.2113)	0.0382(0.2443)	**2.2310sec**(0.0758)	**2660**
Glasso-LDA	**0.7216**(0.1203)	0.2973(0.2039)	0.0155(0.2110)	3.5404sec(0.2109)	113
CLIME-LDA	0.6656(0.1331)	0.3113(0.196)	-0.0208(0.2056)	3.9970sec(0.2002)	113

In summary, there is strong evidence that the BD regularization largely improves the high dimensional practices because it retains the majority of the important information on correlations, i.e., significantly large entries in the covariance matrix. Whereas, the existing methods such as the DC regularization may lead to less efficient statistical inference. This phenomenon confirms that in the extreme scope of ultrahigh dimension and small sample-size, it is almost impossible to explore the exact pairwise correlations of variables. In contrast, using some powerful regularization such as the BD regularization can greatly strengthen the efficiency of statistical inference.

5. Discussion

Generally, there are two options to estimate a high-dimensional precision matrix: a) provide a consistent covariance matrix estimator $\hat{\mathbf{\Sigma}}$ and calculate its inverse; b) provide an estimator of $\mathbf{\Sigma}$ such as the maximum likelihood estimate but estimate $\hat{\mathbf{\Theta}}$ directly from samples through certain iterative algorithms.

The drawbacks of the first option are that $\hat{\mathbf{\Sigma}}$ may be degenerate, the calculation of $\hat{\mathbf{\Sigma}}^{-1}$ becomes instable and time-consuming particularly when $p \gg n$, and storing an ultrahigh dimensional matrix consumes a lot of computer memory. The DC regularization is a frequently-used approach that manages to resolve these drawbacks by simply removing all off-diagonal elements. However, it is an inconsistent estimator of $\mathbf{\Theta}$ and can lead to invalid statistical inferences. The drawbacks of the second option were widely discussed in the literature. For example, in high dimensional setting the iterative algorithm of maximum likelihood estimation takes substantial computing time and large hardware resources in order to find the maximizer.

The proposed BD regularization is in the similar manner with the first option as it is an improvement to the DC regularization. It shares the similar view with the DC regularization in the sense that precise pairwise correlations of X are hardly explored in high-dimensional setting due to limited data information. However, unlike the DC regularization that assumes no correlations among all the components of X, the BD regularization considers that a small number of the components are significantly correlated though the majority of them are not. The resulting estimators of the precision matrix and the covariance matrix are consistent even if $p \gg n$. The innovation of the proposed method is to correctly identify these significantly correlated components, and to shrink the pairwise correlations of all other components to zero. Accordingly, the estimation of a large precision matrix has successfully reduced to a much low-

dimensional one, and the BD regularization is much superior to the DC regularization. Real data analysis in Section 4.2 provides strong evidences that the efficiency of the LDA can be enhanced by accounting for these significantly large pairwise correlations. Otherwise, the LDA discriminator loses the discrimination power if the correlations are ignored completely.

Acknowledgements

We would like to thank the editor and four anonymous reviewers for their very helpful comments and constructive suggestions.

Supplementary Materials

Due to page limitation, more details about the real data and the analysis, the proofs of Theorem 1–3 and three competitor methods of the precision matrix estimation are presented in the Supplementary Materials.

References

Bickel, P.J., Levina, E., 2004. Some theory for fisher's linear discriminant function,naive bayes', and some alternatives when there are many more variables than observations. Bernoulli 10, 989–1010.

Bickel, P.J., Levina, E., 2008a. Covariance regularization by thresholding. The Annals of Statistics 36, 2577–2604.

Bickel, P.J., Levina, E., 2008b. Regularized estimation of large covariance matrices. The Annals of Statistics 36, 199–227.

Cai, T., Liu, W., 2011. Adaptive thresholding for sparse covariance matrix estimation. Journal of the American Statistical Association 106, 672–684.

Cai, T., Liu, W., Luo, X., 2011. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association 106, 594–607.

Candes, E., Tao, T., 2007. The dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics 35, 2313–2351.

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

Fan, J., Kim, D., 2019. Structured volatility matrix estimation for non-synchronized high-frequency financial data. Journal of Econometrics 209, 61–78.

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

Fan, J., Liao, Y., Mincheva, M., 2013. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75, 603–680.

Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 849–911.

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

Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction. volume 2. Springer.

He, K., Kang, J., Hong, H.G., Zhu, J., Li, Y., Lin, H., Xu, H., Li, Y., 2019. Covariance-insured screening. Computational Statistics & Data Analysis 132, 100–114.

Hess, K.R., Anderson, K., Symmans, W.F., Valero, V., Ibrahim, N., Mejia, J.A., Booser, D., Theriault, R.L., Buzdar, A.U., Dempsey, P.J., et al., 2006. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. Journal of Clinical Oncology 24, 4236–4244.

Lauritzen, S.L., 1996. Graphical models. Clarendon Press.

Li, R., Zhong, W., Zhu, L., 2012. Feature screening via distance correlation learning. Journal of the American Statistical Association 107, 1129–1139.

Liu, W., Luo, X., 2015. Fast and adaptive sparse precision matrix estimation in high dimensions. Journal of Multivariate Analysis 135, 153–162.

Mazumder, R., Hastie, T., 2012. Exact covariance thresholding into connected components for large-scale graphical lasso. The Journal of Machine Learning Research 13, 781–794.

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

Page, E.S., 1954. Continuous inspection schemes. Biometrika 41, 100–115.

Pan, W., Wang, X., Xiao, W., Zhu, H., 2019. A generic sure independence screening procedure. Journal of the American Statistical Association 114, 928–937.

Ravikumar, P., Wainwright, M.J., Raskutti, G., Yu, B., 2011. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electronic Journal of Statistics 5, 935–980.

Rothman, A.J., Bickel, P.J., Levina, E., Zhu, J., 2008. Sparse permutation invariant covariance estimation. Electronic Journal of Statistics 2, 494–515.

Rothman, A.J., Levina, E., Zhu, J., 2009. Generalized thresholding of large covariance matrices. Journal of the American Statistical Association 104, 177–186.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58, 267–288.

Witten, D.M., Friedman, J.H., Simon, N., 2011. New insights and faster computations for the graphical lasso. Journal of Computational and Graphical Statistics 20, 892–900.

Yuan, M., 2010. High dimensional inverse covariance matrix estimation via linear programming. The Journal of Machine Learning Research 11, 2261–2286.

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

Zhang, T., Zou, H., 2014. Sparse precision matrix estimation via lasso penalized d-trace loss. Biometrika 101, 103–120.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.