Journal of Statistical Computation and Simulation, 2016
Vol. 86, No. 3, 494–509, http://dx.doi.org/10.1080/00949655.2015.1017823

Tuning-parameter selection in regularized estimations of large covariance matrices

Yixin Fang[a], Binhuan Wang[a*] and Yang Feng[b]

[a]*Division of Biostatistics, New York University School of Medicine, 650 First Avenue, 5th floor, NY 10016, USA;* [b]*Department of Statistics, Columbia University, 1255 Amsterdam Avenue, 10th Floor, NY 10027, USA*

(*Received 26 September 2014; accepted 7 February 2015*)

Recently many regularized estimators of large covariance matrices have been proposed, and the tuning parameters in these estimators are usually selected via cross-validation. However, there is a lack of consensus on the number of folds for conducting cross-validation. One round of cross-validation involves partitioning a sample of data into two complementary subsets, a training set and a validation set. In this manuscript, we demonstrate that if the estimation accuracy is measured in the Frobenius norm, the training set should consist of majority of the data; whereas if the estimation accuracy is measured in the operator norm, the validation set should consist of majority of the data. We also develop methods for selecting tuning parameters based on the bootstrap and compare them with their cross-validation counterparts. We demonstrate that the cross-validation methods with 'optimal' choices of folds are more appropriate than their bootstrap counterparts.

Keywords: banding; bootstrap; covariance matrix; cross-validation; Frobenius norm; operator norm; thresholding

AMS Subject Classification: 62H12

1. Introduction

Estimation of covariance matrices is important in many statistical areas including principal component analysis, linear discriminant analysis, and graphical modelling. Recently, these tools have been used for analysing high-dimensional data sets where the dimensions can be much higher than the sample sizes. Examples include image data, genetic data, and financial data.

Suppose that there are n identically distributed p-dimensional random variables $X_1, \ldots, X_n$ with covariance matrix $\Sigma_{p\times p}$. It is well known that the empirical covariance matrix $\tilde{\Sigma}$ is not a good estimator of Σ when $p > n$, which is defined as follows:

$$\tilde{\Sigma} = [\tilde{\sigma}_{ij}] = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^{\top}, \tag{1}$$

*Corresponding author. Email: binhuan.wang@nyumc.org

where $\tilde{\sigma}_{ij}$ is the element at the ith row and jth column of $\tilde{\Sigma}$, $1 \le i,j \le p$, and $\bar{X} = \sum_{i=1}^{n} X_i/n$. To overcome the curse of dimensionality, many regularized estimators of large covariance matrices have been proposed recently; see [1] and references therein.

1.1. *Two groups of estimators*

There are two main groups of such estimators. One group assumes the covariance matrix being estimated is sparse in the sense that many entries are zero or nearly so. Methods in this group include thresholding [2,3] and generalized thresholding.[4]

Bickel and Levina [2] studied the asymptotic properties of the hard-thresholding estimator,

$$\hat{\Sigma}^{\mathrm{HT}}(\lambda) = [\tilde{\sigma}_{ij} I(|\tilde{\sigma}_{ij}| \ge \lambda)], \tag{2}$$

where $\lambda > 0$ is a tuning parameter to be selected and $I(|\tilde{\sigma}_{ij}| \ge \lambda)$ is an indicator function with value 1 if $|\tilde{\sigma}_{ij}| \ge \lambda$ and 0 otherwise. For the ease of presentation, we use the same notation λ for the tuning parameters in all different kinds of estimators. Rothman et al. [4] proposed a class of thresholding estimators, including the soft-thresholding estimator,

$$\hat{\Sigma}^{\mathrm{ST}}(\lambda) = [\mathrm{sign}(\tilde{\sigma}_{ij})(|\tilde{\sigma}_{ij}| - \lambda)_+], \tag{3}$$

where $(|\tilde{\sigma}_{ij}| - \lambda)_+ = \max(0, |\tilde{\sigma}_{ij}| - \lambda)$.

The other group is for applications where there is a natural metric on the dimensional index set and one expects that the entries farther away from diagonal are smaller. Methods in this group include banding [5,6] and tapering.[7,8]

Bickel and Levina [5] studied the asymptotic properties of the banding estimator,

$$\hat{\Sigma}^{\mathrm{Ba}}(\lambda) = [\tilde{\sigma}_{ij} I(|i-j| \le \lambda)], \tag{4}$$

where integer $0 \le \lambda < p$ is a tuning parameter and $I(|i-j| \le \lambda)$ is an indicator function with value 1 if $|i-j| \le \lambda$ and 0 otherwise. Cai et al. [8] studied the asymptotic properties of the tapering estimator,

$$\hat{\Sigma}^{\mathrm{Ta}}(\lambda) = [w_{ij}^{\lambda} \tilde{\sigma}_{ij}], \tag{5}$$

where w_{ij}^{λ} is a multiplier on each $\tilde{\sigma}_{ij}$ such that for integer $0 \le \lambda < p$, $w_{ij}^{\lambda} = 1$ when $|i-j| \le \lambda/2$, $w_{ij}^{\lambda} = 2 - 2|i-j|/\lambda$ when $\lambda/2 < |i-j| < \lambda$, and $w_{ij}^{\lambda} = 0$ otherwise, $1 \le i,j \le p$.

In this work, we focus on these four estimators, although there are many other methods not belonging to these two groups, such as Cholesky-based regularization [9–11] and factor-based regularization.[12,13]

1.2. *Tuning-parameter selection*

The performance of any estimator depends heavily on the quality of tuning-parameter selection. There are two popular norms which can be used to measure the estimation accuracy, one is the Frobenius norm and the other is the operator norm. For any matrix $M_{p\times p} = [m_{ij}]$, its Frobenius norm and operator norm are defined as

$$\|M\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}^2} \quad \text{and} \quad \|M\|_{\mathrm{op}} = \sup\{\|Mx\|_2 : \|x\|_2 = 1\}, \tag{6}$$

respectively, where $\|\cdot\|_2$ is the Euclidean norm for vectors.

If the estimated Σ were known, for any estimator $\hat{\Sigma}(\lambda)$, we would select the oracle λ as

$$\lambda_F^{\text{oracle}} = \underset{\lambda}{\text{argmin}} \|\hat{\Sigma}(\lambda) - \Sigma\|_F \quad \text{or} \quad \lambda_{\text{op}}^{\text{oracle}} = \underset{\lambda}{\text{argmin}} \|\hat{\Sigma}(\lambda) - \Sigma\|_{\text{op}}, \tag{7}$$

depending on which norm is considered. We call these two tuning methods oracles because they depend on knowing the underlying covariance matrix Σ and they are benchmarks with which those tuning methods described later are compared. In practice, we attempt to estimate the Frobenius risk or the operator risk first,

$$R_F(\lambda) = \mathbb{E}\|\hat{\Sigma}(\lambda) - \Sigma\|_F^2 \quad \text{or} \quad R_{\text{op}}(\lambda) = \mathbb{E}\|\hat{\Sigma}(\lambda) - \Sigma\|_{\text{op}}^2, \tag{8}$$

and then select a value for λ.

The existing theoretical work usually focused on the order of the tuning parameter for a specific model. Xiao and Bunea [14] proved that the banding estimator achieved optimal rate under the operator norm and proposed a Stein's unbiased risk estimate (Sure)-type approach to select the banding parameter, which was of order of n. Li and Zou [15] studied asymptotic properties of SURE information criteria for large covariance matrices with tapering covariance estimator under the Frobenius norm. In this paper, we aim to provide simple and universal guidelines on selecting tuning parameters for different regularized estimators of high-dimensional covariance matrices when a specific type of risk is prioritized (Frobenius risk or operator risk).

The remainder of the manuscript is organized as follows. In Section 2, we describe two popular methods, cross-validation and bootstrap, for estimating the risk under consideration. In Section 3, we conduct extensive simulations to provide some evidences about how many folds for cross-validation would be 'optimal' and whether the 'optimal' cross-validation might be better than the methods based on bootstrap. In Section 4, we illustrate the extent to which the number of folds in cross-validation affects the selected tuning parameter using a real high-dimensional data set. Some conclusions are summarized in Section 5 and the appendix contains all the technical proofs.

2. Methods

2.1. *Cross-validation*

Since the 1970s (e.g. [16]), cross-validation has become one of the most popular methods for tuning-parameter selection. Especially for regularized estimators of large covariance matrices, cross-validation plays a dominant role in tuning-parameter selection. V -fold cross-validation first splits data into $\{\mathcal{D}_1, \ldots, \mathcal{D}_V\}$, and then selects the tuning parameter in $\hat{\Sigma}(\lambda)$ as

$$\lambda_F^{\text{cv}} = \underset{\lambda}{\text{argmin}} \frac{1}{V}\sum_{v=1}^{V} \|\hat{\Sigma}^{(-v)}(\lambda) - \tilde{\Sigma}^{(v)}\|_F^2 \quad \text{or} \quad \lambda_{\text{op}}^{\text{cv}} = \underset{\lambda}{\text{argmin}} \frac{1}{V}\sum_{v=1}^{V} \|\hat{\Sigma}^{(-v)}(\lambda) - \tilde{\Sigma}^{(v)}\|_{\text{op}}^2, \tag{9}$$

where $\tilde{\Sigma}^{(v)}$ is the un-regularized estimator (1) based on $\mathcal{D}_v$ and $\hat{\Sigma}^{(-v)}(\lambda)$ is the regularized estimator under consideration based on data without $\mathcal{D}_v$. Here the size of training data is about $(V-1)n/V$ and the size of validation data is about n/V.

Shao [17] argued that for linear models the cross-validation is asymptotically consistent if the ratio of validation size over sample size goes to one. Motivated by this result, we also consider

reverse cross-validation to select the tuning parameter in $\hat{\Sigma}(\lambda)$ as

$$\lambda_F^{\mathrm{rcv}} = \operatorname*{argmin}_{\lambda} \frac{1}{V}\sum_{v=1}^{V} \|\hat{\Sigma}^{(v)}(\lambda) - \tilde{\Sigma}^{(-v)}\|_F^2 \quad \text{or} \quad \lambda_{\mathrm{op}}^{\mathrm{rcv}} = \operatorname*{argmin}_{\lambda} \frac{1}{V}\sum_{v=1}^{V} \|\hat{\Sigma}^{(v)}(\lambda) - \tilde{\Sigma}^{(-v)}\|_{\mathrm{op}}^2, \tag{10}$$

where $\tilde{\Sigma}^{(-v)}$ is the un-regularized estimator (1) based on data without $\mathcal{D}_v$ and $\hat{\Sigma}^{(v)}(\lambda)$ is the regularized estimator under consideration based on $\mathcal{D}_v$. Here the size of training data is about n/V and the size of validation data is about $(V-1)n/V$. Feng and Yu [18] conducted a systematic study on cross-validation for selecting the optimal tuning parameter in penalized likelihood estimators for generalized linear models.

However, there is a lack of consensus, even discussion, on how many folds should be considered when using cross-validation (or reverse cross-validation) to select tuning parameters in the regularized estimators of large covariance matrices. Here are some examples. In [2], 2-fold cross-validation was used (i.e. the training size is $n_1 = n/2$ and the validation size is $n_2 = n/2$). In [5], reverse 3-fold cross-validation was used (i.e. $n_1 = n/3$ and $n_2 = 2n/3$). In [19,20], 5-fold cross-validation was used (i.e. $n_1 = 4n/5$ and $n_2 = n/5$).

2.2. *Bootstrap*

2.2.1. *Bootstrap for the Frobenius norm.*

Let $\tilde{\Sigma}^s = (n/(n-1))\tilde{\Sigma} = [\tilde{\sigma}_{ij}^s]$ be the usual sample covariance matrix. Let $\hat{\Sigma}(\lambda) = [\hat{\sigma}_{ij}(\lambda)]$ be the regularized estimator under consideration. In their Lemma 1, Yi and Zou [19] showed that the Frobenius risk can be decomposed into

$$\begin{aligned} R_F(\lambda) &= \mathbb{E}\|\hat{\Sigma}(\lambda) - \tilde{\Sigma}^s\|_F^2 + 2\sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{Cov}(\hat{\sigma}_{ij}(\lambda), \tilde{\sigma}_{ij}^s) - \sum_{i=1}^{p}\sum_{j=1}^{p}\mathrm{var}(\tilde{\sigma}_{ij}^s) \\ &= \text{apparent error} + \text{covariance penalty} - \text{constant}, \end{aligned} \tag{11}$$

where terms 'apparent error' and 'covariance penalty' come from [21]. In the same paper, Efron proposed to use the bootstrap method to estimate the covariance penalty. Assume that $\{X_1^{b*}, \ldots, X_n^{b*}\}$, $b = 1, \ldots, B$, are samples repeatedly drawn from some underlying parametric or non-parametric bootstrap model (to be discussed later). For each bootstrap sample, the corresponding estimates $\tilde{\Sigma}^{s,b*} = [\tilde{\sigma}_{ij}^{s,b*}]$ and $\hat{\Sigma}(\lambda)^{b*} = [\hat{\sigma}_{ij}^{b*}(\lambda)]$ are obtained. Then the covariance penalty can be estimated by

$$\widehat{\mathrm{Cov}}(\lambda) = 2\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\frac{1}{B-1}\sum_{b=1}^{B}\hat{\sigma}_{ij}^{b*}(\lambda)\tilde{\sigma}_{ij}^{s,b*} - \frac{1}{B(B-1)}\sum_{b=1}^{B}\hat{\sigma}_{ij}^{b*}(\lambda)\sum_{b=1}^{B}\tilde{\sigma}_{ij}^{s,b*}\right), \tag{12}$$

and the Frobenius risk can be estimated by

$$\hat{R}_F(\lambda) = \|\hat{\Sigma}(\lambda) - \tilde{\Sigma}^s\|_F^2 + \widehat{\mathrm{Cov}}(\lambda), \tag{13}$$

where the constant term in Equation (11) can be ignored for tuning-parameter selection and can be recovered for risk estimation. Then the tuning parameter can be selected as

$$\lambda_F^{\mathrm{boot}} = \operatorname*{argmin}_{\lambda} \hat{R}_F(\lambda). \tag{14}$$

Now we discuss how to select an appropriate bootstrap model for generating bootstrap samples. First, as pointed out by Efron,[21] for high-dimensional applications, parametric bootstrap

is better than non-parametric bootstrap. Second, as pointed out by Efron [21] also, the 'ultimate bigger' bootstrap model,

$$\hat{F} = N(\bar{X}, \tilde{\Sigma}^s), \tag{15}$$

where N stands for multivariate normal distribution, has 'the advantage of not requiring model assumptions', but 'pays for this luxury with increased estimation error'. Third, as discussed in [21], 'the exact choice of $\hat{F}$ is often quite unimportant'. Considering these remarks, in all the numerical results, we consider an intermediate bootstrap model,

$$\hat{F} = N(\bar{X}, \hat{\Sigma}(\hat{\lambda}_0)), \tag{16}$$

where $\hat{\lambda}_0$ is selected via (14) based on the ultimate bootstrap model (15).

2.2.2. *Bootstrap for the operator norm.*

It is very difficult to estimate the operator risk $R_{\text{op}}(\lambda)$, because it cannot be easily decomposed like (11). Here we derive a rough approximation to $R_{\text{op}}(\lambda)$ for banding and tapering estimators and hope this will stimulate more accurate approximations.

For any regularized estimator $\hat{\Sigma}(\lambda)$, let $\Gamma = (\hat{\Sigma}(\lambda) - \Sigma)(\hat{\Sigma}(\lambda) - \Sigma)^\top$ and $\Gamma^* = \mathbb{E}(\Gamma)$. Following the delta-method in [22] and some arguments in the appendix, we have

$$R_{\text{op}}(\lambda) = \mathbb{E}\left(\max_{\|\beta\|_2=1} \beta^\top \Gamma \beta\right) \doteq \max_{\|\beta\|_2=1} \beta^\top \mathbb{E}(\Gamma)\beta + \beta_1^{*\top} \mathbb{E}(\Delta \Pi \Delta)\beta_1^*, \tag{17}$$

where $\Delta = \Gamma - \Gamma^*$ and $\Pi = \sum_{j=2}^p (1/(l_1^* - l_j^*))\beta_j^* \beta_j^{*\top}$ with $\{(\beta_j^*, l_j^*),\ j = 1, \ldots, p\}$ being the eigenvectors and eigenvalues from eigen-system $\Gamma^* \beta = l\beta$. The last term in Equation (17) is known as *Hadamard second variation formula* (e.g. [23]). The approximation still holds if Γ^* is replaced by some unbiased estimator $\hat{\Gamma}^*$; that is,

$$R_{\text{op}}(\lambda) \doteq \hat{l}_1^* + \hat{\beta}_1^{*\top} \mathbb{E}(\hat{\Delta} \hat{\Pi} \hat{\Delta}) \hat{\beta}_1^*, \tag{18}$$

where $\hat{\Delta} = \Gamma - \hat{\Gamma}^*$ and $\hat{\Pi} = \sum_{j=2}^p (\hat{l}_1^* - \hat{l}_j^*)^{-1} \hat{\beta}_j^* \hat{\beta}_j^{*\top}$ with $\{(\hat{\beta}_j^*, \hat{l}_j^*), j = 1, \ldots, p\}$ from eigen-system $\hat{\Gamma}^* \beta = l\beta$. Furthermore, we can estimate the expectation in the second term of Equation (18) via the bootstrap using the same model as Equation (16).

Remark 1 For banding estimator (4) and tapering estimator (5), we derive an unbiased estimator $\hat{\Gamma}^*$ in the appendix. Unfortunately, we fail to derive any unbiased estimator for thresholding estimators (2) and (3).

Remark 2 Based on our limited numerical expericence, the approximation in Equation (17) is very accurate, but due to the curse of dimensionality, the approximation in Equation (18) is rough for high-dimensional data.

3. Simulation results

The data are generated from $N(0, \Sigma)$ with three covariance models adopted from [19] and one model adopted from [24] are considered, sample size n is set as 200, and three settings of dimension are considered, $p = 100$, 200, and 1000.

Model 1. The covariance matrix $\Sigma = [\sigma_{ij}]$, where $\sigma_{ii} = 1$ for $1 \le i \le p$ and $\sigma_{ij} = \rho |i - j|^{-(\alpha+1)}$ for $1 \le i \ne j \le p$. Let $\rho = 0.6$ and $\alpha = 0.1$ or 0.5.

Model 2. The covariance matrix $\Sigma = [\sigma_{ij}]$, where $\sigma_{ij} = \rho^{|i-j|}$ for any $1 \le i,j \le p$. Let $\rho = 0.9$ or 0.5.

Model 3. This model is a truncated version of model 1, where $\sigma_{ii} = 1$ for $1 \le i \le p$ and $\sigma_{ij} = \rho|i-j|^{-(\alpha+1)}I(|i-j| \le 6)$ for $1 \le i \ne j \le p$. Let $\rho = 0.6$ and $\alpha = 0.1$ or 0.5.

Model 4. This model is designed for thresholding with no banding structure in terms of sparsity, where $\sigma_{ij} = s_{ij} \cdot (s_{ii}s_{jj})^{-1/2}$, where $S = I_{p\times p} + U^\top U = (s_{ij})_{p\times p}$ with U being a sparse matrix with κ non-zero entries equal to $+1$ or -1 with equal change. Here we set $\kappa = p$.

Ten cross-validation methods for tuning-parameter selection are compared: 2-fold, 3-fold, 5-fold, 10-fold, 15-fold, and 20-fold cross-validations (CV2, CV3, CV5, CV10, CV15, CV20), 2-fold cross-validation based on 50 random splits (RCV2), reverse 3-fold, 5-fold, and 10-fold cross-validations (reCV3, reCV5, reCV10). The bootstrap methods (bootstrap) are also compared.

The cross-validation using $n_1 = n - \lceil n/\log(n)\rceil$ for training and $n_2 = \lceil n/\log(n)\rceil$ for validation is also implemented, but the results are not reported because this method does not perform very well compared with others although some nice asymptotic property was derived in [2]. The leave-one-out cross-validation is not considered because it is impossible to estimate a covariance matrix using only one data point.

We use the Frobenius norm and the operator norm as evaluation criteria with those four regularized estimators (banding, tapering, hard thresholding and soft thresholding). Each simulation setting is repeated $K = 200$ times, and the performance is measured by the empirical mean square error (MSE), which is the average of 200 values of $\|\hat{\Sigma}(\hat{\lambda}) - \Sigma\|_F^2$ or $\|\hat{\Sigma}(\hat{\lambda}) - \Sigma\|_{\text{op}}^2$.

3.1. *Results in the Frobenius norm*

First, the 10 different cross-validation methods are compared using the Frobenius norm, with results summarized in Figures 1–4. Out-of-chart MSE values are excluded from the figures. Since all the true covariance matrices have some banding or tapering structure, both the banding estimator and the tapering estimator are more accurate than the thresholding estimators.

From Figures 1–4, we see that 10-fold cross-validation performs best for all four models and all four regularized estimators. We also see that 15-fold and 20-fold cross-validations perform comparably with 10-fold cross-validation, but they require more extensive computations. This finding is quite similar to the one in [25], which also suggested 10-fold cross-validation for linear models.

Then 10-fold cross-validation method is compared with the bootstrap method in Section 2.2.1 and the SURE method in [19], with results summarized in Figure 4. Note that the comparison for Model 4 is not shown here because Model 4 is designed for thresholding methods. In [19], the SURE method was compared with 5-fold cross-validation and it was found that the SURE method performs slightly better than 5-fold cross-validation. However, from Figure 5, we see that 10-fold cross-validation performs slightly better than the SURE method. This is consistent with Figures 1–4 in which the 10-fold cross-validation performs slightly better that 5-fold cross-validation. Also note that the SURE method is only applicable for the banding and tapering estimators.

From Figure 5, we also see that the bootstrap method performs very similar to 10-fold cross-validation for the banding and tapering estimators. However, the comparison is complicated for thresholding estimators, because sometimes 10-fold cross-validation performs much better than the bootstrap method whereas sometimes the bootstrap method performs slightly better than 10-fold cross-validation.

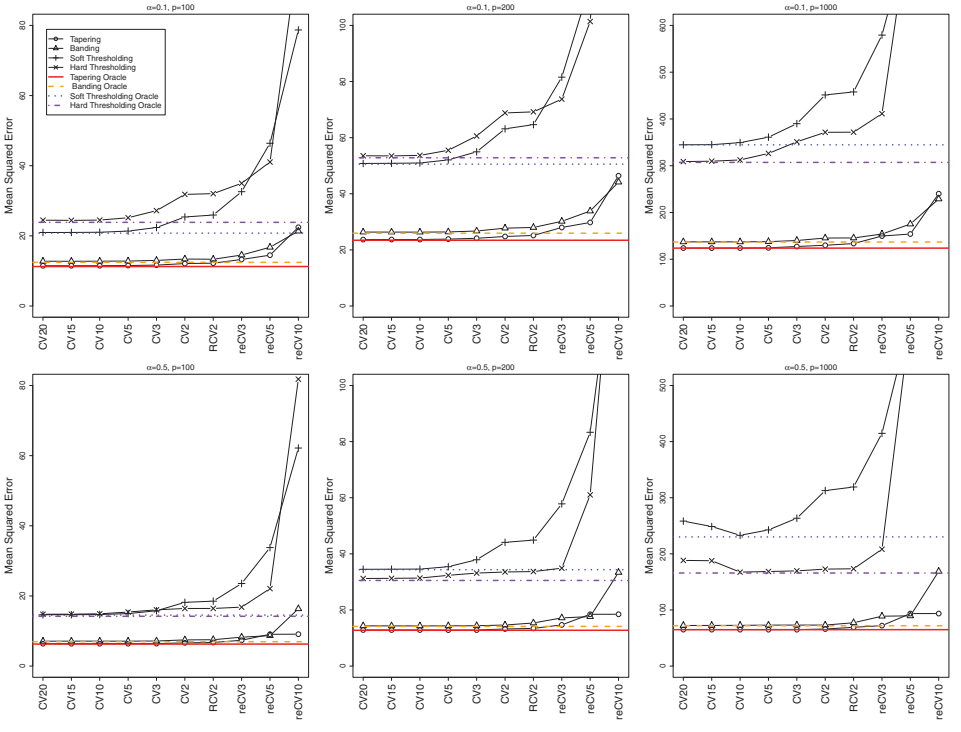

Figure 1. Ten cross-validation methods are compared for Model 1 ($n = 200$; $p = 100, 200, 1000$; $\alpha = 0.1, 0.5$). Performances are measured by MSE in Frobenius norm.

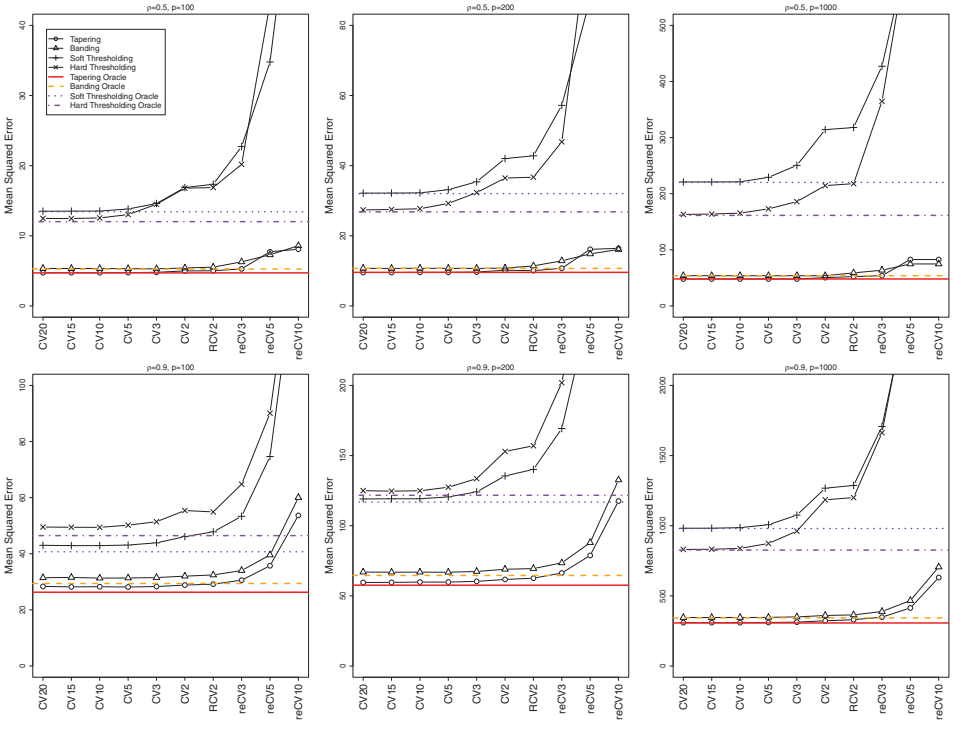

Figure 2. Ten cross-validation methods are compared for Model 2 ($n = 200$; $p = 100, 200, 1000$; $\rho = 0.9, 0.5$). Performances are measured by MSE in Frobenius norm.

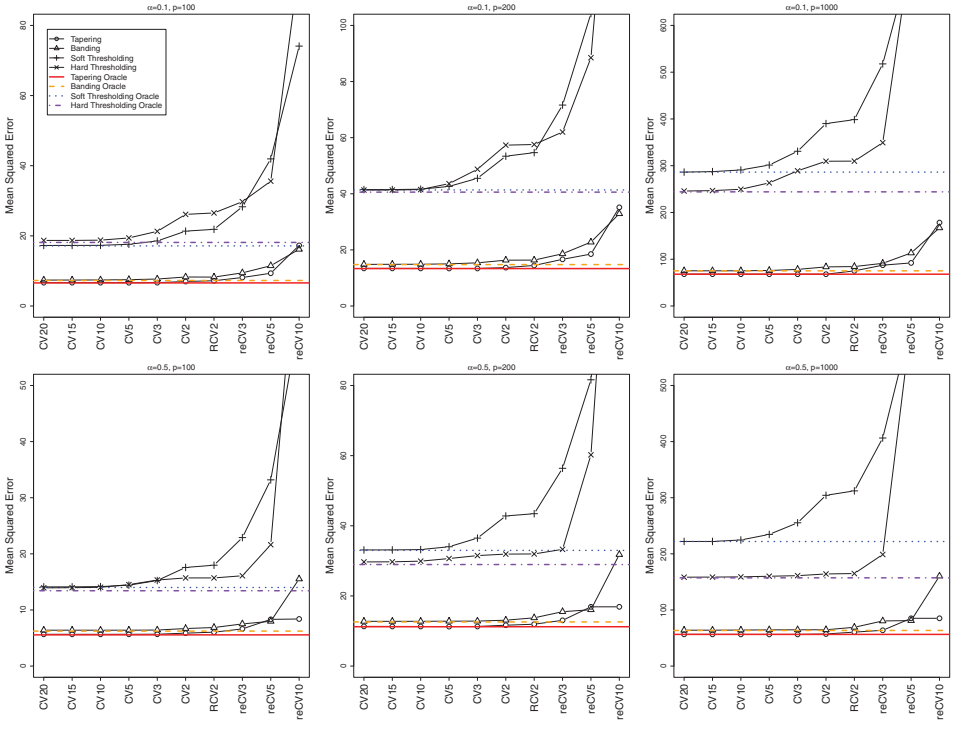

Figure 3. Ten cross-validation methods are compared for Model 3 ($n = 200$; $p = 100, 200, 1000$; $\alpha = 0.1, 0.5$). Performances are measured by MSE in Frobenius norm.

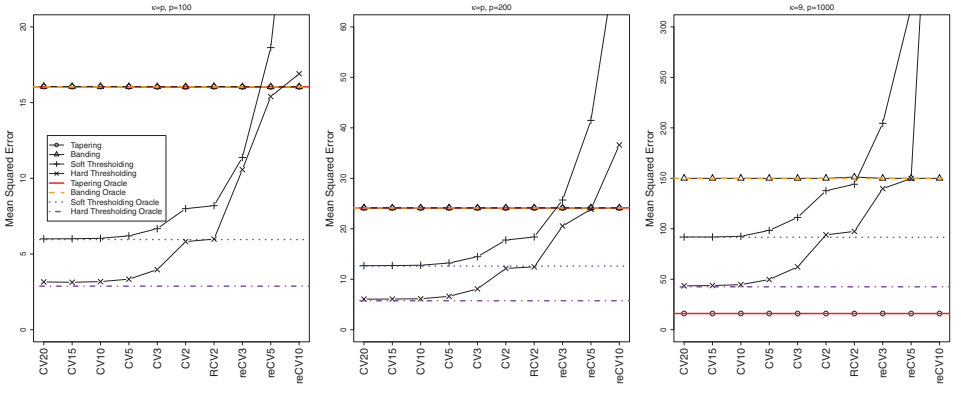

Figure 4. Ten cross-validation methods are compared for Model 4 ($n = 200$; $p = 100, 200, 1000$; $\kappa = p$). Performances are measured by MSE in Frobenius norm.

3.2. *Results in the operator norm*

Again, 10 cross-validation methods are compared using the operator norm, with results summarized in Figures 6–9. We see that, for the banding and tapering estimators, reverse 3-fold cross-validation performs best in most cases, while in other cases it performs almost as well as reverse 5-fold cross-validation. For the hard-thresholding estimator, 2-fold cross-validation or 2-fold cross-validation based on 50 random splits performs the best in all cases. For the

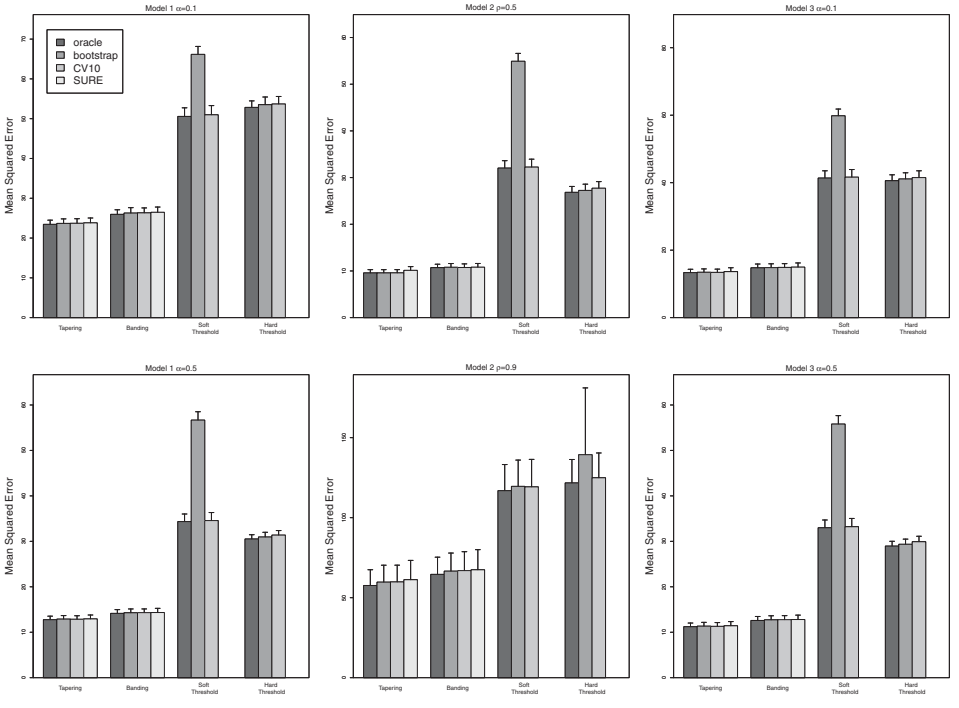

Figure 5. Ten-fold cross-validation is compared with the bootstrap and SURE for Model 1–3 ($n = 200$; $p = 200$). Performances are measured by MSE in Frobenius norm.

soft-thresholding estimator, either 2-fold cross-validation or reverse 3-fold cross-validation performs best in all cases. In addition, it seems using multiple random splits does not improve the performance significantly.

Therefore, from Figure 6–9, we see that either 2-fold cross-validation or reverse 3-fold cross-validation performs best if the MSE is in terms of the operator norm. This finding is different from the result that 10-fold cross-validation performs best for the Frobenius norm. In other words, we need larger training size for the Frobenius norm whereas larger validation size is needed for the operator norm.

For the banding and the tapering estimators, reverse 3-fold cross-validation method is compared with the bootstrap method in Section 2.2.2 for Model 1–3, with results summarized in Figure 10. We see that the bootstrap methods performs similarly as the reverse 3-fold cross-validation for the banding and tapering estimators. The comparison shows that the 'rough' approximation (18) is working well. On the other hand, since the bootstrap does not outperform reverse 3-fold cross-validation and it is much more computationally expensive, we recommend the reverse 3-fold cross-validation over the bootstrap method when the operator norm is considered.

The different performances of cross-validation methods when the target is Frobenius norm or operator norm are very interesting. Intuitively speaking, when minimizing Frobenius norm, we essentially minimize the sum of the squared element-wise estimation error. Note that a counterpart for high-dimensional regression models is the tuning-parameter selection problem when one would like to minimize the L_2 loss for penalized likelihood estimators. When minimizing the operator norm, on the other hand, one needs to impose a higher penalty level since the goal is the spectral behaviour of the matrix estimation error. This is consistent with the well-known

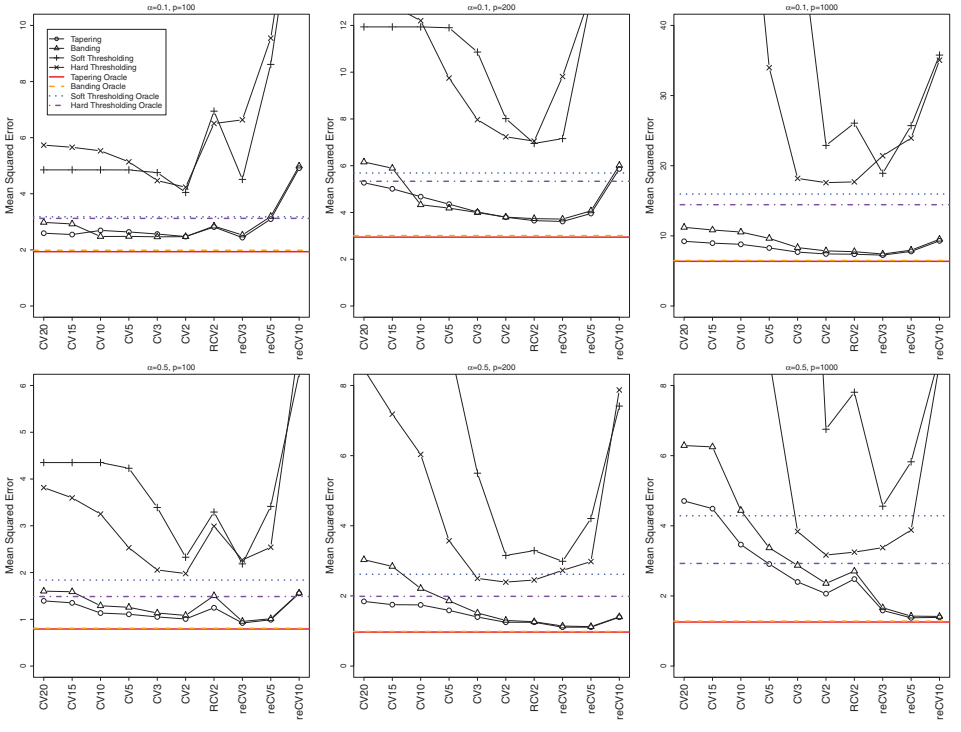

Figure 6. Ten cross-validation methods are compared for Model 1 ($n = 200$; $p = 100, 200, 1000$; $\alpha = 0.1, 0.5$). Performances are measured by MSE in operator norm.

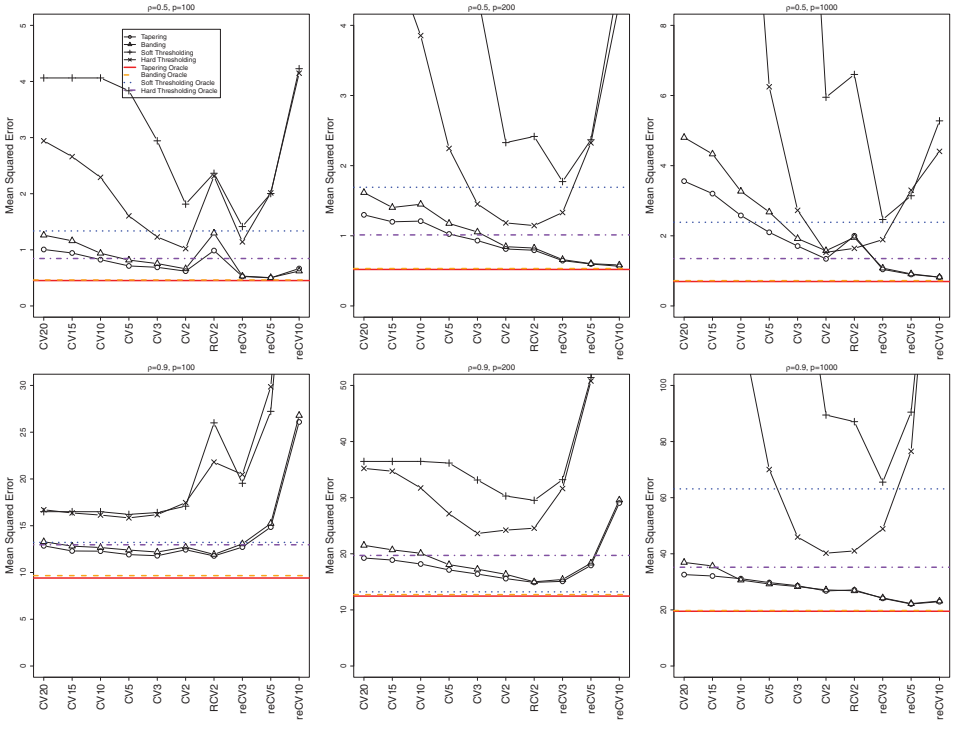

Figure 7. Ten cross-validation methods are compared for Model 2 ($n = 200$; $p = 100, 200, 1000$; $\rho = 0.9; \rho = 0.5$). Performances are measured by MSE in operator norm.

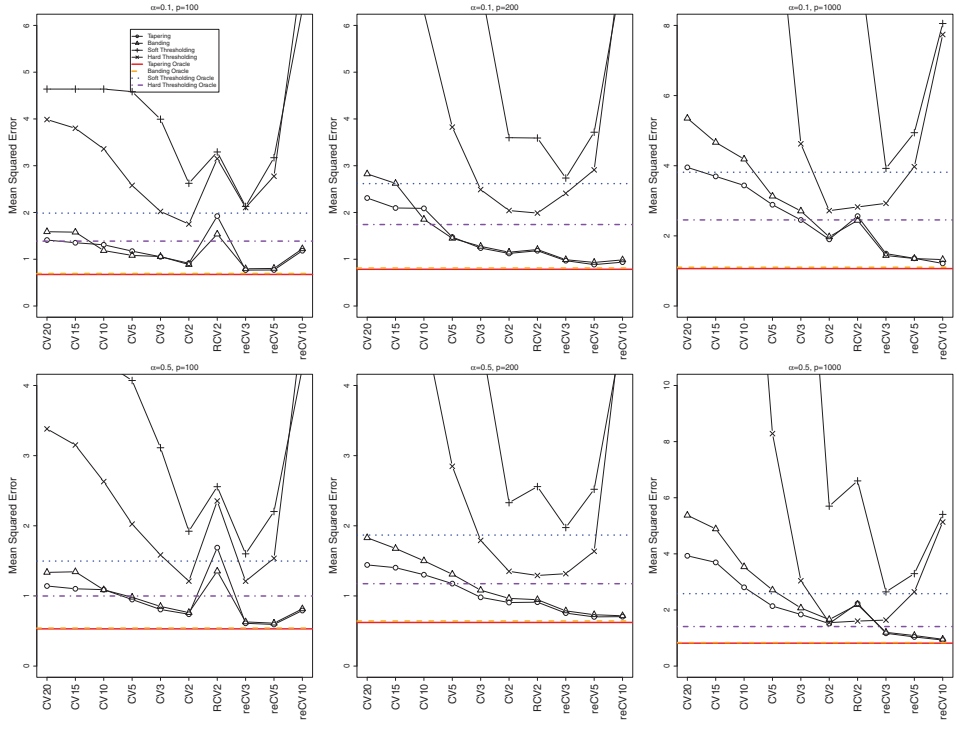

Figure 8. Ten cross-validation methods are compared for Model 3 ($n = 200$; $p = 100, 200, 1000$; $\alpha = 0.1, 0.5$). Performances are measured by MSE in operator norm.

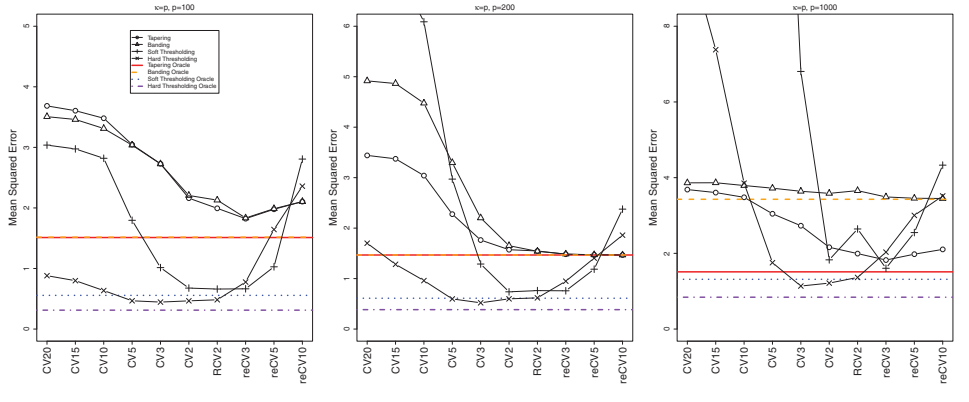

Figure 9. Ten cross-validation methods are compared for Model 4 ($n = 200$; $p = 100, 200, 1000$; $\kappa = p$). Performances are measured by MSE in operator norm.

fact that CV with less fold (or reverse cross-validation) will incur more penalties on the model complexity.

4. Read data analyses

To illustrate the extent to which the number of folds in cross-validation affects the selected tuning parameter, we apply we apply the four regularization methods (banding, tapering,

hard-thresholding and soft-thresholding) to estimate the covariance matrix using a real high-dimensional data set.

The data set we consider is a part of a big data set from a study of metastasis of breast cancer. The full data set contains gene expression, annotations and clinical data, and has been analysed by many investigators including Wang et al. [26] and Minn et al.[27] We use the data set 'vdx' incorporated in R package 'genefu', which consists of 150 subjects with 914 gene-expression variables. The goal is to estimate the 914×914 covariance matrix. The order of variables play an important role in the application of banding and tapering methods and it should be decided according to the nature of the data. For the purpose of illustration, we simply assume the order in the data set is the as the order of gene-expression variables. Note that the order of variables does not matter in the application of hard-thresholding and soft-thresholding methods.

We consider seven kinds of cross-validation with different choices of the number of folds: CV2, CV3, CV5, CV10, CV15, reCV3, and reCV5. Each kind of cross-validation applied to each regularization method leads to a possibly different tuning parameter. Figure 11 displays these selected tuning parameters and it indicates that different cross-validation methods may select significantly different tuning parameters.

From Figure 11, we see that, if the Frobenius norm is considered, the selection of tuning parameter is insensitive to cross-validation methods with majority training data (i.e. CV2, CV3, CV5, CV10, and CV15). Also, for tapering method and soft-thresholding method, the tuning parameters selected by these cross-validation methods are significantly different from the ones selected by cross-validation methods with majority validation data (i.e. reCV3 and reCV5). From Figure 11, we also see that, if the operator norm is considered, the selection of tuning parameter is sensitive to cross-validation methods with different number of folds. Therefore, it suggests

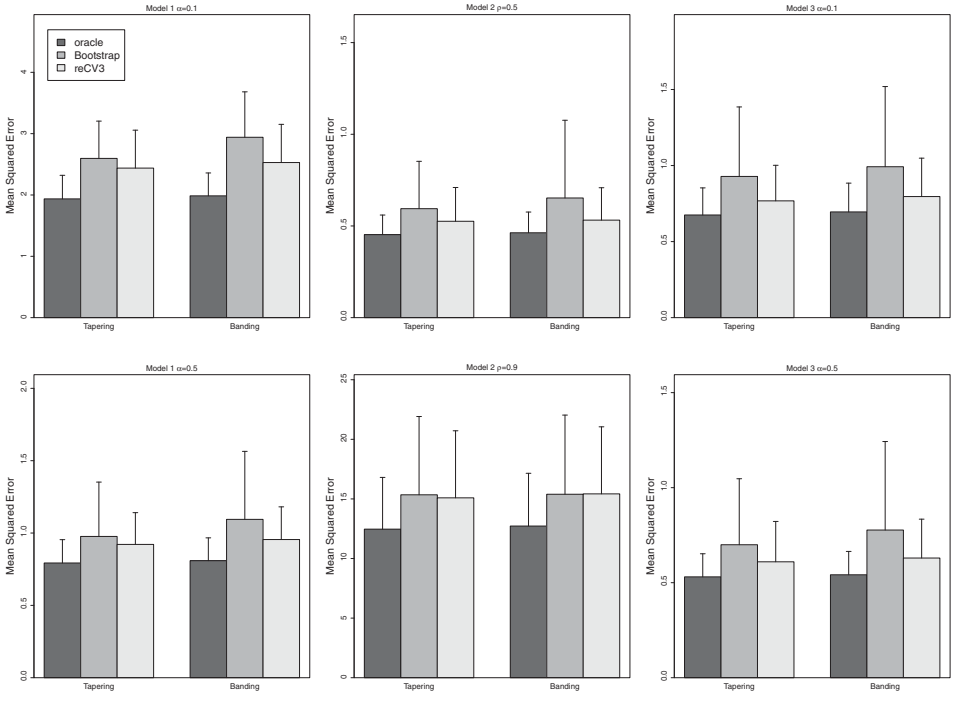

Figure 10. Reverse three-fold cross-validation is compared with the bootstrap method for Model 1–3 ($n = 200$; $p = 200$). Performances are measured by MSE in operator norm.

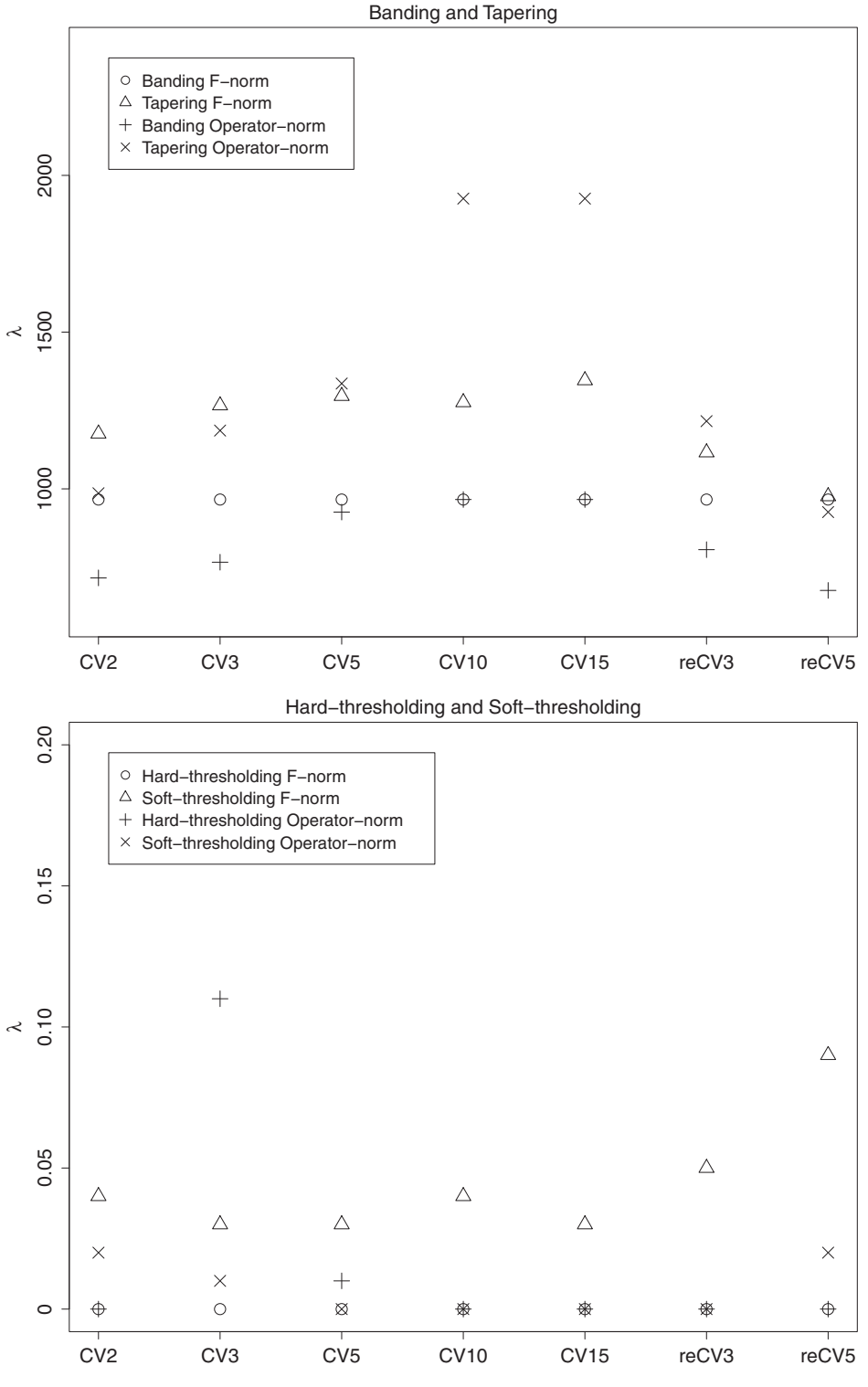

Figure 11. Seven cross-validation methods are compared using the gene-expression data.

that we should be more cautious about tuning parameter selection when the operator norm is considered.

5. Conclusions

In this manuscript, we compare two classes of methods (cross-validation and bootstrap) for selecting tuning parameters in two groups of regularized estimators (banding and thresholding) for covariance matrix, where estimation accuracy is measured in two norms (Frobenius norm and operator norm). Based on extensive simulations and a real data study, we draw the following conclusions:

(1) Cross-validation is computationally convenient and performs better than the methods based on bootstrap;
(2) If the Frobenius norm is considered, we suggest 10-fold cross-validation for both groups of regularized estimators;
(3) If the operator norm is considered, we suggest the validation set should consist of majority of the data.

An R package *CVTuningCov* for implementing the two tuning-parameter selection methods is available on the CRAN website.

Acknowledgments

We would like to thank the editor, the associate editor and two referees for their valuable comments which led to substantial improvements in this paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Dr Yang Feng's work is supported in part by US NSF grant [DMS-1308566].

ORCID

Yang Feng http://orcid.org/0000-0001-7746-7598

References

[1] Johnstone I. On the distribution of the largest eigenvalue in principal components analysis. Ann Statist. 2001;29:295–327.
[2] Bickel P, Levina E. Covariance regularization by thresholding. Ann Statist. 2008;36:2577–2604.
[3] El Karoui N. Operator norm consistent estimation of large-dimensional sparse covariance matrices. Ann Statist. 2008;36:2717–2756.
[4] Rothman AJ, Levina E, Zhu J. Generalized thresholding of large covariance matrices. J Amer Statist Assoc. 2009;104:177–186.
[5] Bickel PJ, Levina E. Regularized estimation of large covariance matrices. Ann Statist. 2008;36:199–227.
[6] Wu W, Pourahmadi M. Banding sample autocovariance matrices of stationary processes. Statist Sinica. 2009;19:1755–1768.
[7] Furrer R, Bengtsson T. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. J Multivariate Anal. 2007;98:227–255.

[8] Cai TT, Zhang C-H, Zhou HH. Optimal rates of convergence for covariance matrix estimation. Ann Statist. 2010;38:2118–2144.
[9] Huang JZ, Liu N, Pourahmadi M, Liu L. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika. 2006;93:85–98.
[10] Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrix estimation. Ann Statist. 2007;37:4254–4278.
[11] Rothman AJ, Levina E, Zhu J. A new approach to Cholesky-based covariance regularization in high dimensions. Biometrika. 2010;97:539–550.
[12] Fan J, Fan Y, Lv J. High dimensional covariance matrix estimation using a factor model. J Econometrics. 2008;147:186–197.
[13] Fan J, Liao Y, Micheva M. Large covariance estimation by thresholding principal orthogonal complements. J R Stat Soc Ser B. 2013;75:603–680.
[14] Xiao L, Bunea F. On the theoretic and practical merits of the banding estimator for large covariance matrices. 2014. Available from: arXiv:1402.0844 [math.ST].
[15] Li D, Zou H. SURE information criteria for large covariance matrix estimation and their asymptotic properties. 2014. Available from: arXiv:1406.6514 [math.ST].
[16] Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc Ser B. 1974;36:111–147.
[17] Shao J. Linear model selection by cross-validation. J Amer Statist Assoc. 1993;88:486–494.
[18] Feng Y, Yu Y. Consistent cross-validation for tuning parameter selection in high-dimensional variable selection. 2013. Available from: arXiv:1308.5390 [stat.ME].
[19] Yi F, Zou H. SURE-tuned tapering estimation of large covariance matrices. Comput Statist Data Anal. 2013;58:339–351.
[20] Xue L, Ma S, Zou H. Positive definite l_1 penalized estimation of large covariance matrices. J Amer Statist Assoc. 2012;107:1480–1491.
[21] Efron B. The estimation of prediction error: covariance penalties and cross-validation. J Amer Statist Assoc. 2004;99:619–632.
[22] Silverman BW. Smoothed functional principal components analysis by choice of norm. Ann Statist. 1996;24:1–24.
[23] Tao T. Topics in random matrix theory, Graduate Studies in Mathematics, Vol. 132. Providence (RI): American Mathematical Society; 2012.
[24] Cai TT, Zhou HH. Minimax estimation of large covariance matrices under l_1 norm (with discussion). Statist Sinica. 2012;22:1319–1378.
[25] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 2, 1995 August 20–25; Montréal, Québec, Canada. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 1995. p. 1137–1143.
[26] Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005;365:671–679.
[27] Minn AJ, Gupta GP, Padua D, Bos P, Nguyen DX, Nuyten D, Kreike B, Zhang Y, Wang Y, Ishwaran H, Foekens JA, van de Vijver M, Massague J. Lung metastasis genes couple breast tumor size and metastatic spread. Proc Natl Acad Sci. 2007;104:6740–6745.

Appendix

A.1. *Approximation in Equation* (17)

Let $\{(\beta_j, l_j),\ j = 1, \ldots, p\}$ be eigenvectors and eigenvalues from eigen-system $\Gamma\beta = l\beta$. Following the delta method used in [22], let

$$\Gamma = \Gamma^* + \varepsilon\Delta^0,$$
$$\beta_1 = \beta_1^* + \varepsilon\beta^{*(1)} + \alpha\beta^{*(2)} + \ldots,$$
$$l_1 = l_1^* + \varepsilon l^{*(1)} + \alpha l^{*(2)} + \ldots,$$

where $\varepsilon = n^{-1/2}$ and $\alpha = n^{-1}$. Because $\Gamma\beta_1 = l_1\beta$, we have

$$(\Gamma^* + \varepsilon\Delta^0)(\beta_1^* + \varepsilon\beta^{*(1)} + \alpha\beta^{*(2)} + \ldots) = (l_1^* + \varepsilon l^{*(1)} + \alpha l^{*(2)} + \ldots)(\beta_1^* + \varepsilon\beta^{*(1)} + \alpha\beta^{*(2)} + \ldots).$$

Comparing the coefficients of powers of $\ddot{I}_{\dagger}$ and α on both sides of this equation, we have

$$l_1 \doteq l_1^* + \varepsilon\beta_1^{*\top}\Delta^0\beta_1^* + \alpha\beta_1^{*\top}\Delta^0\Pi\Delta^0\beta_1^*,$$

whose expectation is Equation (17).

A.2. *Unbiased estimator of* Γ

With a slight abuse of notation, for both the banding estimator and the tapering estimator, let

$$\hat{\Sigma}(\lambda) = [w_{ij}^{\lambda}\tilde{\sigma}_{ij}^{s}].$$

Let $W^{\lambda} = [w_{ij}^{\lambda}]$, whose jth column is defined as w_j^{λ}. Also let Σ_j be the jth column of Σ and let $W_j^{\lambda} = \mathrm{diag}(w_j^{\lambda})$. Note that $(n-1)\tilde{\Sigma}^s \sim W_p(n-1, \Sigma)$, where W stands for Wishart distribution. By some tedious arguments, we have

$$\Gamma^* = \frac{1}{n-1}\sum_{j=1}^{p} W_j^{\lambda}[\sigma_{jj}\Sigma + \Sigma_j\Sigma_j^{\top}]W_j^{\lambda} + \sum_{j=1}^{p}(W_j^{\lambda} - I_p)\Sigma_j\Sigma_j^{\top}(W_j^{\lambda} - I_p).$$

In order to find an unbiased estimator for Σ^*, it suffices to find an unbiased estimator for $\sigma_{kl}\sigma_{k'l'}$ for any $1 \le k, l, k', l' \le p$. Let $X_i = (X_{i1}, \ldots, X_{ip})^{\top}$, $\bar{X}_j = \sum_i X_{ij}/n$, and $\bar{X}_j^{(-i)} = \sum_{i' \ne i} X_{i'j}/(n-1)$. In this manuscript, we use

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_{ik} - \bar{X}_k)(X_{il} - \bar{X}_l) \times \frac{1}{n-2}\sum_{i' \ne i}(X_{i'k'} - \bar{X}_{k'}^{(-i)})(X_{i'l'} - \bar{X}_{l'}^{(-i)})$$

to estimate $\sigma_{kl}\sigma_{k'l'}$.