Computational Statistics and Data Analysis 58 (2013) 339–351

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

SURE-tuned tapering estimation of large covariance matrices

Feng Yi, Hui Zou [*]

School of Statistics, University of Minnesota, Minneapolis, MN 55455, United States

ARTICLE INFO

Article history:
Received 23 July 2011
Received in revised form 7 February 2012
Accepted 12 September 2012
Available online 23 September 2012

Keywords:
Covariance matrix
Cross-validation
Frobenius norm
Operator norms
SURE
Tapering estimator

ABSTRACT

Bandable covariance matrices are often used to model the dependence structure of variables that follow a nature order. It has been shown that the tapering covariance estimator attains the optimal minimax rates of convergence for estimating large bandable covariance matrices. The estimation risk critically depends on the choice of the tapering parameter. We develop a Stein's Unbiased Risk Estimation (SURE) theory for estimating the Frobenius risk of the tapering estimator. SURE tuning selects the minimizer of SURE curve as the chosen tapering parameter. An extensive Monte Carlo study shows that SURE tuning is often comparable to the oracle tuning and outperforms cross-validation. We further illustrate SURE tuning using rock sonar spectrum data. The real data analysis results are consistent with simulation findings.

1. Introduction

Suppose we observe independent and identically distributed p-dimensional random variables $X_1, \ldots, X_n$ with covariance matrix $\Sigma_{p\times p}$. The usual sample covariance matrix is an excellent estimator for $\Sigma_{p\times p}$ in the conventional setting where p is small and fixed and the sample size n diverges to infinity. Nowadays, massive high-dimensional data are more and more common in scientific investigations, such as imaging, web mining, microarrays, risk management, spatial and temporal data, and so on. In high-dimensional settings, the sample covariance matrix performs very poorly; see Johnstone (2001) and references therein. To overcome the difficulty imposed by high dimensions, many regularized estimates of large covariance matrices have been proposed in the recent literature. These regularization methods include Cholesky-based penalization (Huang et al., 2006; Lam and Fan, 2007; Rothman et al., 2010), thresholding (Bickel and Levina, 2008a; El Karoui, 2008; Rothman et al., 2009), banding (Bickel and Levina, 2008b; Wu and Pourahmadi, 2009) and tapering (Furrer and Bengtsson, 2007; Cai et al., 2010). In particular, the tapering estimator is shown to be minimax rate optimal for estimating the bandable covariance matrices that are often used to model the dependence structure of variables that follow a nature order (Cai et al., 2010; Cai and Zhou, 2010). Much of the published theoretical work assumes the data follow a normal distribution, although some have relaxed the normality assumption to a tail probability condition such as sub-Gaussian distribution assumption. Nevertheless, the lower bound results in the minimax estimation theory were actually established for a family of multivariate normal distributions (Cai et al., 2010; Cai and Zhou, 2010). In this paper, we consider the tapering estimator under the normal distribution assumption.

We begin with some notation and definitions. Let $\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$ denote the Frobenius norm of A. Let $\|A\|_q$ denote the ℓ_q operator norm of A. When $q = 1$, the ℓ_1 norm is $\max_i \sum_j |a_{ij}|$; when $q = 2$, the ℓ_2 norm is equal to the largest singular

[*] Corresponding author.
E-mail addresses: fengyi@stat.umn.edu (F. Yi), hzou@stat.umn.edu (H. Zou).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.09.007

value of A. Consider the following parameter spaces:

$$\mathcal{F}_\alpha = \left\{ \Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \le Mk^{-\alpha} \text{ for all } k, \text{ and } \lambda_{\max}(\Sigma) \le M_0 \right\},$$
$$\mathcal{F}'_\alpha = \left\{ \Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \le Mk^{-\alpha} \text{ for all } k, \text{ and } \max_i \sigma_{ii} \le M_0 \right\},$$

where α, M, M_0 are positive constants. The parameter α specifies the rate of decay of the off-diagonal elements of Σ as they move away from the diagonal. A larger α parameter indicates a higher degree of "sparsity". Thus we can also regard α as a *sparsity index* of the parameter space. Let $\tilde{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T - \bar{X}\bar{X}^T$ be the MLE of Σ. The tapering estimator (Cai et al., 2010) is defined as

$$\breve{\Sigma}^{(k)} = (\breve{\sigma}_{ij}^{(k)})_{1\le i,j\le p} = (w_{ij}^{(k)}\tilde{\sigma}_{ij})_{1\le i,j\le p},$$

where, for a tapering parameter k,

$$w_{ij}^{(k)} = \begin{cases} 1, & \text{when } |i-j| \le k/2 \\ 2 - \dfrac{|i-j|}{k/2}, & \text{when } k/2 < |i-j| < k \\ 0, & \text{otherwise.} \end{cases} \tag{1.1}$$

Tapering is a generalization of banding where $\hat{\sigma}_{ij}^{B(k)} = I(|i-j| \le k)\tilde{\sigma}_{ij}$. We assume $p \ge n$ and $\log(p) = o(n)$ in the sequel. We cite the following results (Cai et al., 2010; Cai and Zhou, 2010):

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha} p^{-1}\mathbb{E}\|\hat{\Sigma} - \Sigma\|_F^2 \asymp n^{-(2\alpha+1)/(2\alpha+2)}, \tag{1.2}$$

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}_\alpha} \mathbb{E}\|\hat{\Sigma} - \Sigma\|_2^2 \asymp n^{-2\alpha/(2\alpha+1)} + \frac{\log(p)}{n}, \tag{1.3}$$

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}'_\alpha} \mathbb{E}\|\hat{\Sigma} - \Sigma\|_1^2 \asymp n^{-\alpha/(\alpha+1)} + \frac{\log(p)}{n}, \tag{1.4}$$

where $a_n \asymp b_n$ if there are positive constants c_1 and c_2 independent of n such that $c_1 \le a_n/b_n \le c_2$. Furthermore, define three tapering parameters as following

$$\begin{aligned} &k_F = n^{1/(2\alpha+2)}, \qquad k_2 = n^{1/(2\alpha+1)} \\ &k_1 = \min\{n^{1/(2\alpha+2)}, (n/\log(p))^{1/(2\alpha+1)}\}. \end{aligned} \tag{1.5}$$

Then the tapering estimator with $k = k_F$, $k = k_2$ and $k = k_1$ attains the minimax bound in (1.2)–(1.4), respectively.

The minimax rate optimal choices of k shed light on the importance of choosing the right tapering parameter. However, there are at least two difficulties in using the minimax theory to construct the tapering parameter. First, the minimax tapering estimators depend on α. If α is unknown, which is often the case in reality, then the minimax optimal tapering "estimators" are not real estimators. Second, the minimax rate optimal tapering estimators can be conservative for estimating some covariance matrices. For instance, assume that the data are generated from a normal distribution with a $MA(1)$ covariance where $\sigma_{ij} = I(i=j) + 0.5I(|i-j| = 1)$. Although this covariance matrix is in $\mathcal{F}_\alpha$ for $\alpha > 0$, the optimal k should be 2 no matter which matrix norm is used. Therefore, it is desirable to have a reliable data-driven method to choose the tapering parameter. Tuning is usually done by first constructing an estimate of the risk for each k and then picking the minimizer of the estimated risk curve. Cross-validation and Bootstrap are the popular nonparametric techniques for that purpose. Bickel and Levina (2008a,b) discussed the use of two-fold cross-validation for selecting the banding parameter of the banding estimator. They claimed that although cross-validation estimates the risk very poorly, it can still select the banding parameter quite well.

In this paper, we suggest a different tuning method by borrowing the idea in Stein's unbiased risk estimation (SURE) theory (Stein, 1981; Efron, 1986, 2004). Compared with cross-validation, the SURE approach is computationally less expensive and provides a much better estimate of the Frobenius risk. The explicit form of SURE formula is derived in Section 2. Here we demonstrate the effectiveness of SURE tuning in Fig. 1 where we compare the true Frobenius risk curve (as a function of k) and the SURE curves. We generated the data from the simulation model used in Cai et al. (2010). Two α values were used: $\alpha = 0.1$ corresponds to a dense covariance model and $\alpha = 0.5$ corresponds to a sparse covariance model. Fig. 1 clearly shows three important points. First, the average of 100 SURE curves is virtually identical to the Frobenius risk curve, which agrees with the SURE theory as shown in Section 2. Second, the minimizer of each SURE curve is very close to the minimizer of the true risk curve. Third, the minimizer of each cross-validation curve is also close to the minimizer of the

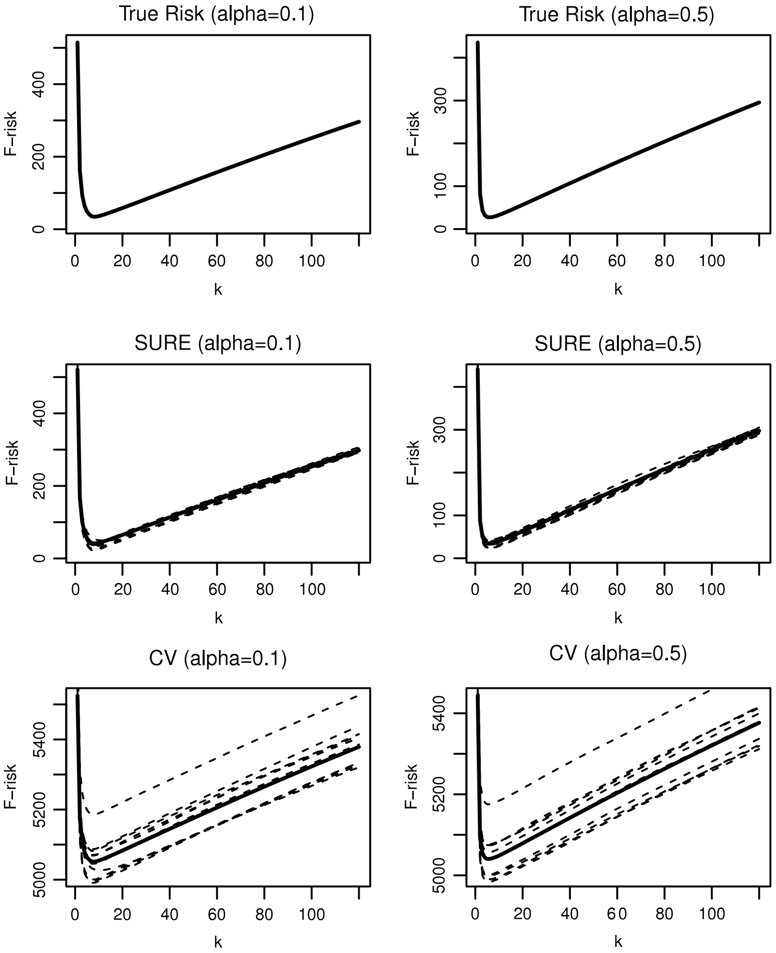

Fig. 1. Comparing the true risk curve, the SURE curve and the CV curve under the Frobenius norm. The data are generated from the simulation model 1 in Section 3 with $n = 250$, $p = 500$, $\alpha = 0.1$ and 0.5. In the second row we plot 10 SURE curves (dashed lines) and the average of 100 SURE curves (the solid line). Similar plots are shown in the third row for cross-validation.

true risk curve, but the cross-validation estimator of the Frobenius risk is way too large. The true risk is within [100, 500] while the cross-validation risk is within [5000, 5500]. In practice we not only want to select a good model but also want to understand how well the model performs. Efron (2004) did a careful comparison between SURE and cross-validation and concluded that with minimal modeling SURE can significantly outperform cross-validation. Fig. 1 suggests that Efron's conclusion continues to hold in the covariance matrix estimation problem.

2. Stein's unbiased risk estimation in covariance matrix estimation

In this section, we develop a SURE theory for estimating the Frobenius risk of a weighted MLE, denoted by $\widehat{\Sigma}^{(k)}$, which has the expression $\widehat{\Sigma}_{ij}^{(k)} = w_{i,j}^{(k)}\tilde{\sigma}_{ij}$ where $w_{i,j}^{(k)}$ only depends on i, j, k. The tapering and banding estimators are special examples of the weighted MLE. Tapering weights are defined in (1.1). The banding estimator (Bickel and Levina, 2008b) uses simpler weights $w_{i,j}^{(k)} = I(|i-j| \leq k)$.

The basic idea in SURE can be traced back to the James–Stein estimator of multivariate normal mean. Efron (1986, 2004) studied the use of SURE in estimating prediction error and he named it covariance penalty method. Shen and Ye (2002) applied the covariance penalty idea to perform adaptive model selection. Donoho and Johnstone (1995) developed SureShrink for adaptive wavelet thresholding. Efron et al. (2004) and Zou et al. (2007) applied SURE to Lasso model selection.

2.1. SURE identity

For an arbitrary estimator $\widehat{\Sigma}$ of the covariance matrix, the Frobenius risk $(\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_F^2)$ is equivalent to the squared ℓ_2 risk for estimating the vector $(\sigma_{11}, \ldots, \sigma_{1p}, \ldots, \sigma_{p1}, \ldots, \sigma_{pp})^T$. As the first step of SURE, we derive a covariance penalty identity for the matrix Frobenius risk of an arbitrary estimator of Σ.

Lemma 1. *Let $\tilde{\Sigma}^s = \frac{n}{n-1}\tilde{\Sigma}$ be the usual sample covariance matrix. For an arbitrary estimator of Σ, denoted by $\widehat{\Sigma} = (\hat{\sigma}_{ij})$, its Frobenius risk can be written as*

$$\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_F^2 = \mathbb{E}\|\widehat{\Sigma} - \tilde{\Sigma}^s\|_F^2 - \sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{var}(\tilde{\sigma}_{ij}^s) + 2\sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{cov}(\hat{\sigma}_{ij}, \tilde{\sigma}_{ij}^s). \tag{2.1}$$

The second term in the right hand of (2.1) is the same for all estimators of Σ. Thus, if we only care of comparing the Frobenius risk of different estimators, the second term can be dropped and we can write

$$\begin{aligned} PR(\widehat{\Sigma}) &= \mathbb{E}\|\widehat{\Sigma} - \tilde{\Sigma}^s\|_F^2 + 2\sum_{i=1}^{p}\sum_{j=1}^{p}\operatorname{cov}(\hat{\sigma}_{ij}, \tilde{\sigma}_{ij}^s) \\ &= \text{Apparent error} + \text{Optimism}, \end{aligned} \tag{2.2}$$

where *PR* stands for prediction risk and we have borrowed Efron's terminology 'apparent error' and 'optimism' (Efron, 2004). The optimism is expressed by a covariance penalty term. Since $\|\widehat{\Sigma} - \tilde{\Sigma}^s\|_F^2$ is an automatic unbiased estimate of the apparent error, it suffices to construct a good estimate of the optimism in order to estimate *PR*.

For the weighted MLE, we observe that $\operatorname{cov}(\hat{\sigma}_{ij}^{(k)}, \tilde{\sigma}_{ij}^s) = w_{ij}^{(k)}\frac{n-1}{n}\operatorname{var}(\tilde{\sigma}_{ij}^s)$. The next lemma provides a nice unbiased estimator of $\operatorname{var}(\tilde{\sigma}_{ij}^s)$.

Lemma 2. *If $\{X_i\}_{i=1}^n$ is a random sample from $N(\mu, \Sigma)$, then*

$$\operatorname{var}(\tilde{\sigma}_{ij}^s) = \frac{\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}}{n-1}, \tag{2.3}$$

and an unbiased estimate of $\operatorname{var}(\tilde{\sigma}_{ij}^s)$ is given by $\widehat{\operatorname{var}}(\tilde{\sigma}_{ij}^s)$ which equals

$$\frac{n^2(n^2-n-4)}{(n-1)^2(n^3+n^2-2n-4)}\tilde{\sigma}_{ij}^2 + \frac{n^3}{(n-1)(n^3+n^2-2n-4)}\tilde{\sigma}_{ii}\tilde{\sigma}_{jj}. \tag{2.4}$$

From (2.3) we see the MLE for $\operatorname{var}(\tilde{\sigma}_{ij}^s)$ is $\frac{\tilde{\sigma}_{ij}^2 + \tilde{\sigma}_{ii}\tilde{\sigma}_{jj}}{n-1}$, which is almost identical to the unbiased estimator in (2.4). We prefer to use an exact unbiased estimate of the optimism. In addition, the unbiased estimator in (2.4) is the UMVUE of $\operatorname{var}(\tilde{\sigma}_{ij}^s)$.

Lemma 2 shows that an unbiased estimator for $PR(\widehat{\Sigma}^{(k)})$ is given by

$$\widehat{PR}(k) = \|\widehat{\Sigma}^{(k)} - \tilde{\Sigma}^s\|_F^2 + \sum_{1\le i,j\le p}\left(2w_{ij}^{(k)}\frac{n-1}{n}\right)\widehat{\operatorname{var}}(\tilde{\sigma}_{ij}^s). \tag{2.5}$$

Similarly, an unbiased estimator for $\mathbb{E}\|\widehat{\Sigma}^{(k)} - \Sigma\|_F^2$ is given by

$$\begin{aligned} \text{SURE}(k) &= \|\widehat{\Sigma}^{(k)} - \tilde{\Sigma}^s\|_F^2 + \sum_{1\le i,j\le p}\left(2w_{ij}^{(k)}\frac{n-1}{n} - 1\right)\widehat{\operatorname{var}}(\tilde{\sigma}_{ij}^s) \\ &= \sum_{1\le i,j\le p}\left(\frac{n}{n-1} - w_{ij}^{(k)}\right)^2\tilde{\sigma}_{ij}^2 + \sum_{1\le i,j\le p}\left(2w_{ij}^{(k)} - \frac{n}{n-1}\right)(a_n\tilde{\sigma}_{ij}^2 + b_n\tilde{\sigma}_{ii}\tilde{\sigma}_{jj}) \end{aligned} \tag{2.6}$$

with $a_n = \frac{n(n^2-n-4)}{(n-1)(n^3+n^2-2n-4)}$ and $b_n = \frac{n^2}{n^3+n^2-2n-4}$.

2.2. SURE tuning

Once the tapering estimator is constructed, the SURE formula automatically provides a good estimate of its Frobenius risk. Naturally we use $\hat{k}^{\text{sure}}$ as the tapering parameter under the Frobenius norm where

$$\hat{k}^{\text{sure}} = \arg\min_k \text{SURE}(k). \tag{2.7}$$

Unfortunately we do not have a direct SURE formula for the matrix ℓ_q norm, $q = 1, 2$. We suggest using $\hat{k}^{\text{sure}}$ as the tapering parameter for both ℓ_1 and ℓ_2 norms as well. We list several good reasons for using this selection strategy.

1. One can expect the optimal tapering parameter should be the same under different matrix norm if the underlying covariance matrix is an exactly banded matrix, i.e., there is a constant k_0 such that $\sigma_{ij} = 0$ whenever $|i-j| > k_0$. Hence, it is reasonable to expect that the optimal choices of the tapering parameter under the Frobenius norm and the matrix ℓ_1, ℓ_2 norms stay close if the underlying covariance model is very sparse.

2. Cai and Zhou (2010) showed that as long as $\log(p) \le n^{1/(2\alpha+2)}$, the minimax optimal tapering parameters under the ℓ_1 norm and the Frobenius norm are the same. This can be easily seen from (1.5).
3. The ℓ_2 norm is the most popular matrix operator norm. We argue that minimizing the Frobenius norm leads to a good estimator, although may not be the best, under the ℓ_2 norm. From Cai et al. (2010) we know that

$$\sup_{\mathcal{F}_\alpha} \mathbb{E}\|\check{\Sigma}^{(k)} - \Sigma\|_2^2 \le C\left[k^{-2\alpha} + \frac{k+\log(p)}{n}\right] \equiv C \cdot R_2(k).$$

Letting $k = k_F = n^{1/(2\alpha+2)}$ yields

$$R_2(k_F) = O(n^{-\alpha/(\alpha+1)} + \log(p)/n).$$

Compare the rate to the minimax optimal rate $n^{-2\alpha/(2\alpha+1)} + \log(p)/n$.
4. As shown in simulation, SURE selection is very stable, although it is biased under the ℓ_1, ℓ_2 norms. Selection stability is a very important concern in model selection (Breiman, 1996). In contrast, even the oracle tuning under the ℓ_1, ℓ_2 norms can show very high variability when the underlying covariance matrix is not very sparse.

3. Monte Carlo study

In this section, we conduct extensive simulation to compare SURE tuning with cross-validation and oracle tuning.

3.1. Models and tuning methods

The data are generated from $N(0, \Sigma)$. Six covariance models are considered.

Model 1. This model is adopted from Cai et al. (2010). The covariance matrix has the form

$$\sigma_{ij} = \begin{cases} 1, & 1 \le i = j \le p \\ \rho|i-j|^{-(\alpha+1)} & 1 \le i \ne j \le p. \end{cases}$$

We let $\rho = 0.6$, $\alpha = 0.1, 0.5$, $n = 250$ and $p = 250, 500, 1000$.

Model 2. The covariance matrix has the form $\sigma_{ij} = \rho^{|i-j|}$, $1 \le i, j \le p$. We let $\rho = 0.95, 0.5$, $n = 250$ and $p = 250, 500, 1000$. This is a commonly used autoregressive covariance matrix for modeling spatial–temporal dependence.

Model 3. This simulation model is a truncated version of model 1. The covariance matrix has the form

$$\sigma_{ij} = \begin{cases} 1, & 1 \le i = j \le p \\ \rho|i-j|^{-(\alpha+1)} I(|i-j| \le 6) & 1 \le i \ne j \le p. \end{cases}$$

We let $\rho = 0.6$, $\alpha = 0.1, 0.5$, $n = 250$ and $p = 250, 500, 1000$. Model 3 represents an exactly banded covariance matrix. It is the sparest among all three simulation models.

Model 4. The covariance matrix has the form

$$\sigma_{ij} = \begin{cases} 1, & 1 \le i = j \le p \\ \rho|i-j|^{-(\alpha+1)}(-1)^{|i-j|} & 1 \le i \ne j \le p. \end{cases}$$

We let $\rho = 0.6$, $\alpha = 0.1, 0.5$, $n = 250$ and $p = 250, 500, 1000$. This model is similar to Model 1 but has negative correlations.

Model 5. σ_{ij} has the form of $\sigma_{ij} = \rho^{|i-j|}(-1)^{|i-j|}$, $1 \le i, j \le p$. We let $\rho = 0.6$, $\alpha = 0.1, 0.5$, $n = 250$ and $p = 250, 500, 1000$. This model is similar to Model 2 but has negative correlations.

Model 6. The covariance matrix has the form

$$\sigma_{ij} = \begin{cases} 1, & 1 \le i = j \le p \\ \rho|i-j|^{-(\alpha+1)} I(|i-j| \le 6)(-1)^{|i-j|} & 1 \le i \ne j \le p. \end{cases}$$

We let $\rho = 0.6$, $\alpha = 0.1, 0.5$, $n = 250$ and $p = 250, 500, 1000$. This model is similar to Model 3 but has negative correlations.

For each covariance model, the theoretical optimal tapering parameters are defined as $k_a^{\text{opt}} = \arg\min_k \mathbb{E}\|\check{\Sigma}^{(k)} - \Sigma\|_a^2$, where $a = F, 1, 2$. In our simulation study the risk curves can be computed numerically, and thus we can find the numerical values of k_a^{opt} for $a = F, 1, 2$.

We considered three tuning techniques in the simulation study: SURE, cross-validation and oracle tuning. The oracle tuning is defined as

$$\hat{k}_a^{\text{oracle}} = \arg\min_k \|\check{\Sigma}^{(k)} - \Sigma\|_a^2$$

Table 1
Simulation model 1: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 1: Tapering parameter selection												
p	α	k^{opt}			$\hat{k}^{\text{oracle}}$			$\hat{k}^{\text{sure}}$	$\hat{k}^{\text{cv}}$		
		F	ℓ_1	ℓ_2	F	ℓ_1	ℓ_2	F, ℓ_1, ℓ_2	F	ℓ_1	ℓ_2
250	0.1	11	9	30	10.70 (0.56)	10.46 (3.03)	36.29 (8.52)	10.63 (1.18)	9.66 (1.02)	18.34 (9.50)	48.97 (27.15)
250	0.5	6	5	9	5.99 (0.41)	5.88 (1.60)	10.56 (2.21)	6.15 (0.73)	5.46 (0.67)	10.28 (6.24)	20.41 (11.8)
500	0.1	11	9	39	10.83 (0.43)	9.96 (2.60)	44.57 (8.37)	10.52 (0.88)	9.35 (0.73)	19.75 (10.40)	50.56 (23.76)
500	0.5	6	5	10	6.04 (0.28)	5.52 (1.72)	10.64 (2.02)	6.11 (0.60)	5.29 (0.46)	12.08 (5.48)	21.08 (11.30)
1000	0.1	11	9	51	10.92 (0.31)	9.60 (2.37)	55.91 (8.02)	10.65 (0.64)	9.22 (0.54)	18.67 (10.09)	70.68 (29.88)
1000	0.5	6	5	10	6.00 (0.14)	5.24 (1.45)	11.03 (1.83)	6.14 (0.47)	5.17 (0.38)	10.74 (5.67)	28.25 (14.88)

where $a = F, 1, 2$. The idea of oracle tuning is intuitive. Suppose that we could use an independent validation data set of size m ($m \geq n$) for tuning. The chosen k is then found by comparing $\widehat{\Sigma}^{(k)}$ and $\tilde{\Sigma}_m$ under a given matrix norm, where $\tilde{\Sigma}_m$ is the MLE of Σ using the independent validation set. Now imagine m could be as large as we wish. The oracle tuning is basically the independent-validation-set tuning with infinitely many data. The oracle tuning is not realistic but serves as a golden benchmark to check the performance of practical tuning methods.

Cross-validation is a commonly-used practical tuning method. Randomly split the training data into V parts. For $v = 1, \ldots, V$, we leave observations in the vth part as validation data and compute a MLE of Σ, denoted by $\tilde{\Sigma}_v$. Let $\breve{\Sigma}^{(k)}_{-v}$ denote the tapering estimator computed on the rest $V - 1$ parts. Then the cross-validation choices of k under the Frobenius norm and the matrix ℓ_1, ℓ_2 norm are defined as $\hat{k}^{cv}_a = \arg\min_k \frac{1}{V} \sum_{v=1}^{V} \| \breve{\Sigma}^{(k)}_{-v} - \tilde{\Sigma}_v \|_a^2$ where $a = F, 1, 2$, denoting the Frobenius, ℓ_1, ℓ_2 norms. Five-fold cross-validation was used in our simulation.

We also considered an unconventional cross-validation called cv-F that always uses Frobenius-norm for tuning even when the ℓ_1 or ℓ_2 norm is used to evaluate the risk of the tapering estimator. Note that cv-F is a direct analogue of SURE tuning. Since CV is good at capturing the shape of Frobenius risk although the magnitude is too large, cv-F is expected to perform similarly to SURE. But cv-F is still computationally more expensive than SURE.

3.2. *Results and conclusions*

For each model we compared the chosen tapering parameters by oracle, SURE and cross-validation to the optimal tapering parameter and compared the estimation risk of the three tuned tapering covariance estimators. Tables 1–12 summarize the simulation results. We have the following remarks.

1. Under the Frobenius norm, SURE works as well as the oracle tuning. Cross-validation is slightly worse than SURE. SURE and cv-F have very similar performance as expected.
2. Cross-validation completely fails under the ℓ_1, ℓ_2 norms. We can understand the failure of cross-validation under the ℓ_1, ℓ_2 norms by looking at its selection variability. Even the oracle tuning exhibits high variability when the covariance matrix is dense. Under the ℓ_1, ℓ_2 norms, SURE and cv-F still perform quite well comparable to the oracle tuning. Note that SURE and cv-F are very stable.
3. The performance of tuning depends on the degree of sparsity of the underlying covariance model. When the covariance matrix is sparse (models 1,4 with $\alpha = 0.5$, models 2,5 with $\rho = 0.5$ and models 3,6), SURE and cv-F are closer to the oracle tuning. This is not surprising because it is relatively easier to estimate a sparse covariance matrix than a dense one.

4. Rock sonar spectrum data

In this section, we use the sonar data to illustrate the efficacy of SURE tuning and to further demonstrate the conclusions made in the simulation study. The sonar data is publicly available from the UCI repository of machine learning databases (Frank and Asuncion, 2010). We consider its subset consisting of 97 sonar spectra bounced off from rocks. Each spectrum has 60 frequency band energy measurements. Although the dimension is 60, this is still a relatively large dimension scenario, because the sample size is 97. We examined the entries of sample covariance matrix and found there is a quite obvious decay pattern as the entries move away from the diagonal. Hence we used tapering to regularize the sample covariance matrix. SURE and cross-validation were used to select the tapering parameter. Bootstrap was used to assess the variability of each tuning procedure.

Table 2
Simulation model 1: Frobenius, ℓ_1 ℓ_2 risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 1: Estimation risk										
	p	α	Oracle		SURE		CV		CV-F	
Frobenius norm	250	0.1	26.04	(0.11)	26.23	(0.11)	26.30	(0.10)	26.30	(0.10)
	250	0.5	13.63	(0.07)	13.77	(0.07)	13.83	(0.07)	13.83	(0.07)
	500	0.1	53.33	(0.14)	53.54	(0.14)	53.82	(0.14)	53.82	(0.14)
	500	0.5	27.48	(0.11)	27.65	(0.11)	27.87	(0.11)	27.87	(0.11)
	1000	0.1	108.11	(0.21)	108.29	(0.22)	109.15	(0.21)	109.15	(0.21)
	1000	0.5	55.03	(0.14)	55.25	(0.14)	55.04	(0.15)	55.04	(0.15)
ℓ_1 norm	250	0.1	14.17	(0.12)	14.78	(0.15)	17.84	(0.50)	14.78	(0.15)
	250	0.5	3.67	(0.05)	3.87	(0.06)	5.22	(0.34)	3.86	(0.05)
	500	0.1	18.94	(0.14)	19.58	(0.17)	24.20	(0.71)	19.51	(0.15)
	500	0.5	4.22	(0.04)	4.43	(0.06)	5.62	(0.22)	4.40	(0.05)
	1000	0.1	24.08	(0.13)	24.88	(0.17)	29.85	(0.88)	24.73	(0.16)
	1000	0.5	4.64	(0.04)	4.87	(0.05)	6.49	(0.24)	4.78	(0.04)
ℓ_2 norm	250	0.1	2.96	(0.05)	5.35	(0.07)	4.29	(0.16)	5.71	(0.07)
	250	0.5	0.88	(0.01)	1.09	(0.02)	1.48	(0.08)	1.19	(0.02)
	500	0.1	4.26	(0.05)	7.87	(0.07)	5.27	(0.16)	8.45	(0.06)
	500	0.5	0.99	(0.01)	1.23	(0.01)	1.59	(0.07)	1.37	(0.01)
	1000	0.1	5.82	(0.05)	10.56	(0.06)	7.36	(0.19)	11.40	(0.05)
	1000	0.5	1.08	(0.01)	1.33	(0.01)	2.09	(0.10)	1.52	(0.01)

Table 3
Simulation model 2: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 2: Tapering parameter selection											
p	ρ	k^{opt}			$\hat{k}^{\mathrm{oracle}}$			$\hat{k}^{\mathrm{sure}}$	$\hat{k}^{\mathrm{cv}}$		
		F	ℓ_1	ℓ_2	F	ℓ_1	ℓ_2	F, ℓ_1, ℓ_2	F	ℓ_1	ℓ_2
250	0.95	71	71	76	70.79 (4.53)	72.84 (11.93)	77.36 (17.32)	71.23 (12.45)	68.64 (12.92)	80.07 (28.30)	88.24 (33.14)
250	0.50	5	5	5	5.00 (0.00)	4.84 (0.93)	5.13 (1.02)	5.03 (0.17)	5.00 (0.00)	7.87 (6.09)	13.18 (11.93)
500	0.95	70	68	69	70.10 (3.08)	69.50 (12.17)	72.51 (17.00)	70.76 (6.14)	68.04 (6.41)	88.77 (30.46)	107.52 (33.82)
500	0.50	5	5	5	5.00 (0.00)	4.89 (0.90)	5.17 (1.00)	5.00 (0.00)	5.00 (0.00)	8.60 (4.55)	16.68 (15.84)
1000	0.95	69	67	71	69.71 (2.16)	69.83 (11.95)	73.83 (11.68)	70.66 (3.86)	67.48 (3.83)	92.29 (30.56)	117.41 (33.84)
1000	0.50	5	5	5	5.00 (0.00)	4.73 (0.93)	5.00 (0.94)	5.00 (0.00)	5.00 (0.00)	8.85 (6.04)	21.08 (20.90)

Table 4
Simulation model 2: Frobenius, ℓ_1, ℓ_2 risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 2: Estimation risk										
	p	α	Oracle		SURE		CV		CV-F	
Frobenius norm	250	0.95	118.09	(2.66)	125.00	(2.88)	126.19	(2.86)	126.19	(2.86)
	250	0.50	9.88	(0.06)	9.91	(0.07)	9.88	(0.06)	9.88	(0.06)
	500	0.95	250.53	(3.54)	256.94	(3.62)	258.10	(3.59)	258.10	(3.59)
	500	0.50	19.10	(0.08)	19.81	(0.08)	19.81	(0.08)	19.81	(0.08)
	1000	0.95	512.13	(4.90)	517.94	(4.92)	519.26	(4.90)	519.26	(4.90)
	1000	0.50	39.72	(0.11)	39.72	(0.11)	39.72	(0.11)	39.72	(0.11)
ℓ_1 norm	250	0.95	142.91	(5.17)	158.36	(5.80)	176.09	(8.29)	159.29	(5.79)
	250	0.50	1.33	(0.03)	1.39	(0.03)	2.29	(0.27)	1.37	(0.03)
	500	0.95	183.55	(5.21)	198.28	(5.97)	233.56	(9.67)	197.97	(5.79)
	500	0.50	1.43	(0.02)	1.46	(0.03)	2.54	(0.17)	1.46	(0.03)
	1000	0.95	210.56	(3.98)	223.65	(4.76)	279.71	(12.01)	222.86	(4.58)
	1000	0.50	1.58	(0.03)	1.64	(0.03)	3.04	(0.33)	1.64	(0.03)
ℓ_2 norm	250	0.95	36.90	(1.61)	42.98	(1.95)	44.87	(2.02)	43.77	(1.98)
	250	0.50	0.47	(0.01)	0.49	(0.01)	0.89	(0.07)	0.49	(0.01)
	500	0.95	47.09	(1.41)	54.45	(2.06)	66.64	(2.96)	54.82	(2.04)
	500	0.50	0.51	(0.01)	0.53	(0.01)	1.18	(0.10)	0.53	(0.01)
	1000	0.95	56.70	(1.40)	62.31	(1.79)	78.59	(2.85)	62.76	(1.80)
	1000	0.50	0.59	(0.01)	0.61	(0.01)	1.58	(0.14)	0.61	(0.01)

Table 5
Simulation model 3: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 3: Tapering parameter selection											
p	α	k^{opt}			$\hat{k}^{\text{oracle}}$			$\hat{k}^{\text{sure}}$	$\hat{k}^{\text{cv}}$		
		F	ℓ_1	ℓ_2	F	ℓ_1	ℓ_2	F, ℓ_1, ℓ_2	F	ℓ_1	ℓ_2
250	0.1	8	7	7	7.91 (0.29)	7.21 (0.77)	7.56 (1.12)	7.93 (0.26)	7.35 (0.48)	11.15 (5.81)	17.19 (12.54)
250	0.5	6	5	5	5.97 (0.41)	5.57 (1.30)	5.91 (1.14)	6.13 (0.68)	5.47 (0.64)	8.76 (4.64)	13.79 (9.34)
500	0.1	8	7	7	8.00 (0.00)	7.06 (0.81)	7.29 (1.09)	7.93 (0.26)	7.22 (0.42)	11.21 (5.87)	19.49 (18.70)
500	0.5	6	5	5	5.97 (0.17)	5.49 (1.10)	5.59 (1.01)	6.18 (0.59)	5.41 (0.59)	9.95 (8.39)	15.39 (10.43)
1000	0.1	8	7	7	8.00 (0.00)	6.77 (0.90)	6.99 (1.12)	8.00 (0.61)	7.12 (0.33)	11.26 (6.10)	21.79 (17.94)
1000	0.5	6	5	5	6.00 (0.00)	5.13 (1.28)	5.31 (1.20)	6.13 (0.37)	5.20 (0.40)	8.96 (5.72)	18.24 (13.66)

Table 6
Simulation model 3: Frobenius, ℓ_1 ℓ_2 risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 3: Estimation risk										
	p	α	Oracle		SURE		CV		CV-F	
Frobenius norm	250	0.1	13.89	(0.09)	13.93	(0.09)	14.09	(0.09)	14.09	(0.09)
	250	0.5	11.63	(0.07)	11.75	(0.07)	11.82	(0.07)	11.82	(0.07)
	500	0.1	27.68	(0.13)	27.73	(0.13)	28.08	(0.13)	28.08	(0.13)
	500	0.5	23.42	(0.10)	23.59	(0.11)	23.78	(0.10)	23.78	(0.10)
	1000	0.1	55.79	(0.22)	55.79	(0.22)	56.68	(0.22)	56.68	(0.22)
	1000	0.5	46.95	(0.16)	47.06	(0.16)	47.70	(0.14)	47.70	(0.14)
ℓ_1 norm	250	0.1	1.98	(0.04)	2.10	(0.04)	3.42	(0.30)	2.05	(0.04)
	250	0.5	1.47	(0.03)	1.60	(0.03)	2.38	(0.18)	1.59	(0.03)
	500	0.1	2.18	(0.04)	2.36	(0.05)	3.79	(0.34)	2.26	(0.04)
	500	0.5	1.65	(0.02)	1.78	(0.03)	3.62	(0.55)	1.75	(0.03)
	1000	0.1	2.49	(0.04)	2.72	(0.05)	4.34	(0.48)	2.55	(0.05)
	1000	0.5	1.88	(0.03)	2.07	(0.05)	3.34	(0.30)	1.98	(0.04)
ℓ_2 norm	250	0.1	0.67	(0.01)	0.72	(0.02)	1.33	(0.09)	0.71	(0.02)
	250	0.5	0.53	(0.01)	0.58	(0.01)	0.94	(0.06)	0.57	(0.01)
	500	0.1	0.78	(0.02)	0.85	(0.02)	1.66	(0.16)	0.82	(0.02)
	500	0.5	0.59	(0.01)	0.63	(0.01)	1.18	(0.08)	0.62	(0.01)
	1000	0.1	0.88	(0.01)	0.98	(0.02)	2.02	(0.14)	0.93	(0.02)
	1000	0.5	0.69	(0.01)	0.76	(0.02)	1.54	(0.10)	0.73	(0.01)

Table 7
Simulation model 4: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 4: Tapering parameter selection											
p	α	k^{opt}			$\hat{k}^{\text{oracle}}$			$\hat{k}^{\text{sure}}$	$\hat{k}^{\text{cv}}$		
		F	ℓ_1	ℓ_2	F	ℓ_1	ℓ_2	F, ℓ_1, ℓ_2	F	ℓ_1	ℓ_2
250	0.1	11	9	31	10.76 (0.55)	10.49 (2.94)	36.88 (8.62)	10.44 (1.21)	9.50 (0.97)	18.03 (9.28)	46.96 (24.06)
250	0.5	6	5	9	5.99 (0.44)	5.63 (1.40)	10.64 (2.29)	6.04 (0.76)	5.44 (0.64)	10.11 (5.86)	20.84 (14.70)
500	0.1	11	9	38	10.78 (0.46)	9.66 (2.29)	44.15 (8.37)	10.47 (0.85)	9.36 (0.70)	18.88 (10.07)	56.91 (24.31)
500	0.5	6	5	10	6.01 (0.22)	5.51 (1.58)	10.76 (2.22)	6.11 (0.63)	5.29 (0.50)	11.35 (6.81)	20.58 (13.10)
1000	0.1	11	9	51	10.92 (0.27)	9.10 (2.73)	56.00 (7.28)	10.79 (0.46)	9.26 (0.57)	19.12 (12.11)	63.46 (31.95)
1000	0.5	6	5	10	6.00 (0.14)	5.20 (1.44)	10.41 (2.03)	6.05 (0.46)	5.19 (0.39)	10.31 (6.04)	27.61 (19.52)

Table 8
Simulation model 4: Frobenius, ℓ_1 ℓ_2 risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 4: Estimation risk										
	p	α	Oracle		SURE		CV		CV-F	
Frobenius norm	250	0.1	26.07	(0.09)	26.28	(0.09)	26.38	(0.10)	26.38	(0.10)
	250	0.5	13.59	(0.07)	13.75	(0.07)	13.80	(0.07)	13.80	(0.07)
	500	0.1	53.36	(0.14)	53.54	(0.15)	53.81	(0.14)	53.81	(0.14)
	500	0.5	27.57	(0.11)	27.76	(0.11)	27.99	(0.11)	27.99	(0.11)
	1000	0.1	108.44	(0.21)	108.51	(0.21)	109.35	(0.20)	109.35	(0.20)
	1000	0.5	55.42	(0.18)	55.63	(0.18)	56.22	(0.17)	56.22	(0.17)
ℓ_1 norm	250	0.1	14.14	(0.10)	14.64	(0.12)	17.62	(0.47)	14.58	(0.11)
	250	0.5	3.59	(0.04)	3.80	(0.05)	4.95	(0.24)	3.76	(0.05)
	500	0.1	18.74	(0.11)	19.35	(0.14)	23.31	(0.63)	19.34	(0.12)
	500	0.5	4.24	(0.05)	4.47	(0.06)	6.38	(0.51)	4.41	(0.06)
	1000	0.1	24.15	(0.13)	24.97	(0.17)	30.44	(1.15)	24.80	(0.16)
	1000	0.5	4.60	(0.04)	4.87	(0.06)	6.31	(0.24)	4.74	(0.04)
ℓ_2 norm	250	0.1	2.98	(0.05)	5.49	(0.07)	4.21	(0.15)	5.84	(0.07)
	250	0.5	0.88	(0.01)	1.11	(0.02)	1.44	(0.09)	1.20	(0.02)
	500	0.1	4.23	(0.05)	7.90	(0.06)	5.55	(0.18)	8.45	(0.06)
	500	0.5	1.01	(0.01)	1.26	(0.01)	1.57	(0.09)	1.39	(0.01)
	1000	0.1	5.66	(0.04)	10.44	(0.05)	7.07	(0.20)	11.34	(0.05)
	1000	0.5	1.10	(0.01)	1.36	(0.01)	2.18	(0.13)	1.52	(0.01)

Table 9
Simulation model 5: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 5: Tapering parameter selection											
p	ρ	k^{opt}			$\hat{k}^{\text{oracle}}$			$\hat{k}^{\text{sure}}$	$\hat{k}^{\text{cv}}$		
		F	ℓ_1	ℓ_2	F	ℓ_1	ℓ_2	F, ℓ_1, ℓ_2	F	ℓ_1	ℓ_2
250	0.95	71	71	76	70.79 (4.53)	72.84 (11.93)	77.36 (17.32)	71.01 (12.38)	68.59 (12.80)	80.93 (28.25)	89.33 (33.80)
250	0.50	5	5	5	5.00 (0.00)	4.99 (0.92)	5.18 (0.97)	5.02 (0.14)	5.00 (0.00)	8.93 (6.76)	12.34 (10.86)
500	0.95	70	70	71	70.39 (3.17)	71.40 (12.76)	74.86 (18.99)	70.32 (7.15)	67.13 (7.23)	87.43 (31.87)	110.37 (39.78)
500	0.50	5	5	5	5.00 (0.00)	4.80 (0.90)	5.11 (1.05)	5.00 (0.00)	5.00 (0.00)	8.97 (4.88)	15.95 (13.79)
1000	0.95	69	68	72	69.87 (2.48)	68.65 (11.11)	75.06 (12.49)	70.31 (4.23)	67.37 (4.42)	90.49 (28.50)	119.22 (38.16)
1000	0.50	5	5	5	5.00 (0.00)	4.65 (0.97)	4.86 (0.92)	5.00 (0.00)	5.00 (0.00)	8.03 (5.65)	19.02 (17.53)

Table 10
Simulation model 5: Frobenius, ℓ_1, ℓ_2 risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 5: Estimation risk										
	p	ρ	Oracle		SURE		CV		CV-F	
Frobenius norm	250	0.95	118.09	(2.66)	124.96	(2.88)	126.19	(2.87)	126.19	(2.87)
	250	0.50	9.92	(0.06)	9.93	(0.06)	9.92	(0.06)	9.92	(0.06)
	500	0.95	247.49	(3.90)	254.18	(4.22)	256.02	(4.17)	256.02	(4.17)
	500	0.50	19.81	(0.08)	19.81	(0.08)	19.81	(0.08)	19.81	(0.08)
	1000	0.95	511.21	(6.22)	519.52	(6.53)	520.79	(6.34)	520.79	(6.34)
	1000	0.50	39.80	(0.12)	39.80	(0.12)	39.80	(0.12)	39.80	(0.12)
ℓ_1 norm	250	0.95	142.91	(5.17)	158.30	(5.80)	174.46	(7.75)	159.24	(5.82)
	250	0.50	1.31	(0.02)	1.36	(0.03)	2.66	(0.33)	1.36	(0.03)
	500	0.95	184.75	(5.36)	201.05	(6.86)	236.85	(10.41)	201.38	(6.72)
	500	0.50	1.62	(0.03)	1.68	(0.03)	2.74	(0.18)	1.50	(0.03)
	1000	0.95	209.75	(4.26)	225.51	(5.81)	275.02	(11.77)	223.53	(5.29)
	1000	0.50	1.62	(0.03)	1.68	(0.03)	2.80	(0.34)	1.68	(0.03)
ℓ_2 norm	250	0.95	36.90	(1.61)	43.01	(1.95)	45.23	(2.05)	43.74	(1.99)
	250	0.50	0.45	(0.01)	0.48	(0.01)	0.83	(0.06)	0.47	(0.01)
	500	0.95	48.20	(1.72)	55.50	(2.33)	68.21	(3.84)	56.20	(2.31)
	500	0.50	0.51	(0.01)	0.54	(0.01)	1.15	(0.08)	0.54	(0.01)
	1000	0.95	57.00	(1.56)	63.66	(2.00)	82.40	(3.70)	63.86	(1.90)
	1000	0.50	0.59	(0.01)	0.62	(0.01)	1.48	(0.11)	0.62	(0.01)

Table 11
Simulation model 6: tapering parameter selection. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 6: Tapering parameter selection											
p	α	k^{opt}			$\hat{k}^{\text{oracle}}$			$\hat{k}^{\text{sure}}$	$\hat{k}^{\text{cv}}$		
		F	ℓ_1	ℓ_2	F	ℓ_1	ℓ_2	F, ℓ_1, ℓ_2	F	ℓ_1	ℓ_2
250	0.1	8	7	7	7.91 (0.29)	7.01 (0.77)	7.57 (1.08)	7.89 (0.31)	7.28 (0.45)	10.78 (7.22)	16.28 (11.39)
250	0.5	6	5	5	5.99 (0.41)	5.59 (1.22)	5.96 (1.37)	5.99 (0.70)	5.34 (0.57)	8.93 (4.90)	14.78 (10.48)
500	0.1	8	7	7	7.97 (0.17)	7.15 (0.86)	7.18 (0.98)	7.92 (0.27)	7.19 (0.39)	10.59 (3.94)	19.79 (16.91)
500	0.5	6	5	5	6.00 (0.25)	5.53 (1.34)	5.64 (1.38)	6.07 (0.62)	5.36 (0.56)	9.50 (7.25)	16.49 (14.40)
1000	0.1	8	7	7	7.99 (0.10)	6.93 (0.88)	6.98 (1.06)	7.99 (0.10)	7.11 (0.31)	11.43 (6.87)	24.50 (20.40)
1000	0.5	6	5	5	5.99 (0.10)	5.13 (1.21)	5.52 (1.19)	6.07 (0.46)	5.22 (0.42)	9.86 (6.15)	20.23 (15.90)

Table 12
Simulation model 6: Frobenius, ℓ_1 ℓ_2 risk. We report the average value of 100 replications. Corresponding standard errors are shown in parentheses.

Model 6: Estimation risk										
	p	α	Oracle		SURE		CV		CV-F	
Frobenius norm	250	0.1	13.89	(0.09)	13.95	(0.09)	14.09	(0.09)	14.09	(0.09)
	250	0.5	11.61	(0.07)	11.76	(0.07)	11.82	(0.07)	11.82	(0.07)
	500	0.1	27.82	(0.14)	27.90	(0.14)	28.25	(0.14)	28.25	(0.14)
	500	0.5	23.35	(0.10)	23.54	(0.10)	23.77	(0.10)	23.77	(0.10)
	1000	0.1	56.08	(0.21)	56.10	(0.21)	56.95	(0.21)	56.95	(0.21)
	1000	0.5	46.96	(0.16)	47.13	(0.17)	47.74	(0.15)	47.74	(0.15)
ℓ_1 norm	250	0.1	1.99	(0.04)	2.13	(0.05)	3.51	(0.43)	2.05	(0.05)
	250	0.5	1.46	(0.03)	1.58	(0.03)	2.46	(0.20)	1.56	(0.03)
	500	0.1	2.18	(0.04)	2.35	(0.05)	3.42	(0.20)	2.26	(0.04)
	500	0.5	1.66	(0.03)	1.79	(0.04)	3.23	(0.45)	1.77	(0.04)
	1000	0.1	2.41	(0.04)	2.64	(0.05)	4.53	(0.48)	2.49	(0.04)
	1000	0.5	1.85	(0.03)	2.03	(0.04)	3.64	(0.35)	1.96	(0.03)
ℓ_2 norm	250	0.1	0.70	(0.02)	0.74	(0.02)	1.25	(0.08)	0.73	(0.02)
	250	0.5	0.53	(0.01)	0.57	(0.01)	0.98	(0.06)	0.56	(0.01)
	500	0.1	0.78	(0.02)	0.84	(0.02)	1.66	(0.14)	0.82	(0.02)
	500	0.5	0.62	(0.01)	0.67	(0.02)	1.24	(0.10)	0.67	(0.01)
	1000	0.1	0.86	(0.01)	0.97	(0.02)	2.17	(0.16)	0.91	(0.02)
	1000	0.5	0.68	(0.01)	0.73	(0.02)	1.61	(0.10)	0.71	(0.01)

In Fig. 2 we plot SURE and cross-validated estimates of the Frobenius risk and also show the bootstrap histogram of the selected tapering parameter by SURE and cross-validation. Some interesting phenomena are evident in the figure. First, the two bootstrap histograms clearly show that SURE tuning is less variable than cross-validation. Second, SURE tuning selected the high peak of the SURE bootstrap histogram but cross-validation selected a left tail value of its bootstrap histogram. Third, the cross-validation estimate of the Frobenius risk is much larger than the SURE estimate.

Fig. 3 shows the cross-validation tuning results under the ℓ_1, ℓ_2 norms. The selected tapering parameters under the ℓ_1, ℓ_2 norms are not very different from those under the Frobenius norm. The significant difference is that cross-validation tuning under the ℓ_1, ℓ_2 norms has much flatter bootstrap histograms, indicating much larger variability in selection.

We also repeated the above analysis on the other subset consisting of 111 sonar spectra bounced off from metal cylinders and the conclusions are basically the same. For the sake of space consideration, we opt to present the analysis results and figures in a technical report version of this paper.

In conclusion, what we have observed in this real data example is consistent with the simulation results.

5. Discussion

There are two important issues in any regularized estimation procedure: (1) how to select the regularization parameter? and (2) how to estimate the accuracy of a regularized estimator? In traditional vector-estimation problems such as nonparametric regression or classification, cross-validation is a routinely used method for answering both questions and perform well in general. Efron (2004) has shown that SURE can be more accurate than cross-validation for estimating the risk of a vector estimator. In this paper, we have found that cross-validation does not perform satisfactorily for tuning the

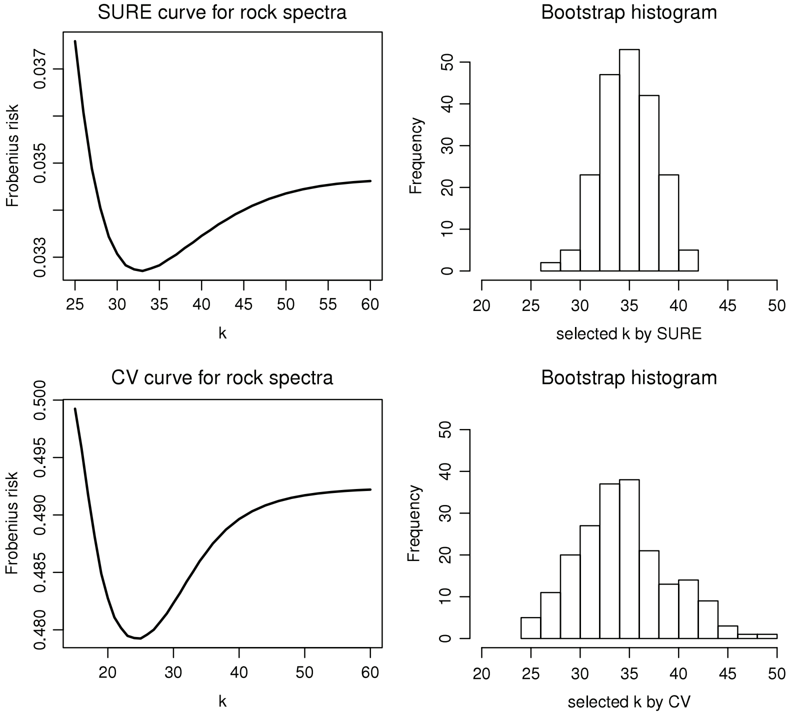

Fig. 2. Rock sonar spectrum data: SURE and cross-validation tuning under the Frobenius norm. The right panels display the bootstrap histograms of the selected tapering parameter by SURE and cross-validation.

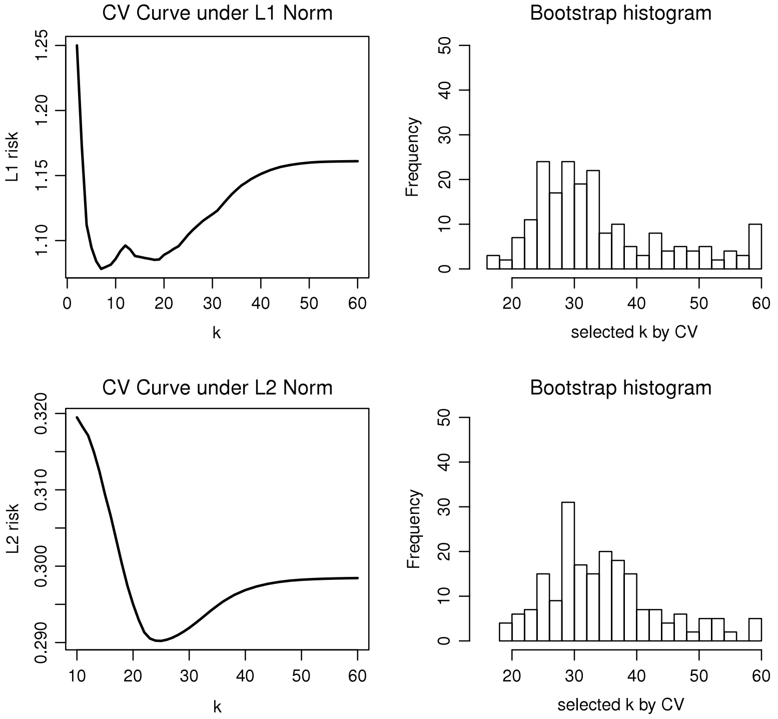

Fig. 3. Rock sonar spectrum data: cross-validation tuning under the ℓ_1, ℓ_2 norms. The right panels display the bootstrap histograms of the selected tapering parameter by cross-validation.

tapering covariance estimator when the objective loss function is the matrix ℓ_1 or ℓ_2 norm. Cross-validation can capture the shape of the Frobenius risk, but the cross-validated estimate of the Frobenius risk tends to be too large to be a good estimate. Our empirical study suggests that the Frobenius norm is better for tuning a covariance matrix estimator even when the objective loss is the ℓ_1 or ℓ_2 norm. To that end, the proposed SURE formula is very useful: it is computationally economic, stable and provides a reliable estimate of the Frobenius risk.

Acknowledgments

This work was supported in part by NSF grant DMS-0846068. The authors thank the editor, AE and referees for their helpful comments.

Appendix

Proof of Lemma 1. We start with Stein's identity (Efron, 2004)

$$(\hat{\sigma}_{ij} - \sigma_{ij})^2 = (\hat{\sigma}_{ij} - \tilde{\sigma}^s_{ij})^2 - (\tilde{\sigma}^s_{ij} - \sigma_{ij})^2 + 2(\hat{\sigma}_{ij} - \sigma_{ij})(\tilde{\sigma}^s_{ij} - \sigma_{ij}). \tag{A.1}$$

Taking expectation at both side of (A.1) and summing over $i, j = 1$ yield

$$\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_F^2 = \mathbb{E}\|\widehat{\Sigma} - \tilde{\Sigma}^s\|_F^2 - \sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{var}(\tilde{\sigma}^s_{ij}) + 2\sum_{i=1}^{p}\sum_{j=1}^{p} \mathrm{cov}(\hat{\sigma}_{ij}, \tilde{\sigma}^s_{ij}).$$

Note that $\mathbb{E}[(\hat{\sigma}_{ij} - \sigma_{ij})(\tilde{\sigma}^s_{ij} - \sigma_{ij})] = \mathrm{cov}(\hat{\sigma}_{ij}, \tilde{\sigma}^s_{ij})$ because $\mathbb{E}\tilde{\sigma}^s_{ij} = \sigma_{ij}$. □

Proof of Lemma 2. The estimators under consideration are translational invariant. Without loss of generality, we can let $\mu = \mathbb{E}(x) = 0$. By straightforward calculation based on bivariate normal distribution, we have

$$\mathbb{E}(x_i^2 x_j^2) = \sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2, \tag{A.2}$$

which holds for both $i = j$ and $i \neq j$.

$$\begin{aligned}\mathbb{E}((\tilde{\sigma}^s_{ij})^2) &= \mathbb{E}\left((n-1)^{-2}\left(\sum_{k=1}^{n} x_{k,i}x_{k,j} - n\bar{x}_i\bar{x}_j\right)^2\right)\\ &= (n-1)^{-2}\left\{\mathbb{E}\left(\left(\sum_{k=1}^{n} x_{k,i}x_{k,j}\right)^2\right) - 2n^{-1}\sum_{k=1}^{n}\mathbb{E}(n\bar{x}_i n\bar{x}_j x_{k,i}x_{k,j}) + n^2\mathbb{E}(\bar{x}_i^2\bar{x}_j^2)\right\}.\end{aligned} \tag{A.3}$$

We also have

$$\begin{aligned}\mathbb{E}\left(\left(n^{-1}\sum_{k=1}^{n} x_{k,i}x_{k,j}\right)^2\right) &= \frac{1}{n}\mathrm{var}(x_i x_j) + (\mathbb{E}(x_i x_j))^2\\ &= \frac{1}{n}(\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2 - \sigma_{ij}^2) + \sigma_{ij}^2\\ &= \frac{1}{n}\sigma_{ii}\sigma_{jj} + \frac{1+n}{n}\sigma_{ij}^2.\end{aligned} \tag{A.4}$$

Note that $\bar{X} \sim N(0, \Sigma/n)$. Using (A.2) we have

$$n^2\mathbb{E}(\bar{x}_i^2\bar{x}_j^2) = 2\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}. \tag{A.5}$$

$$\begin{aligned}\mathbb{E}(n\bar{x}_i n\bar{x}_j x_{k,i}x_{k,j}) &= \sum_{1\le l,l'\le n}\left\{I(l = l' \neq k)\mathbb{E}(x_{l,i}x_{l,j}x_{k,i}x_{k,j}) + I(l = l' = k)\mathbb{E}(x_{k,i}^2 x_{k,j}^2)\right\}\\ &= (n-1)\sigma_{12}^2 + (\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2).\end{aligned} \tag{A.6}$$

Substituting (A.4)–(A.6) into (A.3) gives

$$\mathbb{E}((\tilde{\sigma}^s_{ij})^2) = \frac{n\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}}{n-1}. \tag{A.7}$$

Thus, $\mathrm{var}(\tilde{\sigma}^s_{ij}) = \mathbb{E}((\tilde{\sigma}^s_{ij})^2) - \sigma_{ij}^2 = \frac{\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}}{n-1}$.

We now show (2.4) by deriving an expression for $\mathbb{E}(\tilde{\sigma}^s_{ii}\tilde{\sigma}^s_{jj})$.

$$(n-1)^2\mathbb{E}(\tilde{\sigma}^s_{ii}\tilde{\sigma}^s_{jj}) = \sum_{1\le k,k'\le n}\mathbb{E}(x_{k,i}^2 x_{k',j}^2) - \sum_{1\le k'\le n}\mathbb{E}(\bar{x}_i^2 x_{k',j}^2) - \sum_{1\le k\le n}\mathbb{E}(\bar{x}_j^2 x_{k,i}^2) + n^2\mathbb{E}(\bar{x}_i^2\bar{x}_j^2). \tag{A.8}$$

Repeatedly using (A.2) we have

$$\sum_{1\le k,k'\le n} \mathbb{E}(x_{k,i}^2 x_{k',j}^2) = n^2\sigma_{ii}\sigma_{jj} + 2n\sigma_{ij}^2, \tag{A.9}$$

$$\begin{aligned} n^2\mathbb{E}(\bar{x}_i^2 x_{k',j}^2) &= \sum_{1\le l,l'\le n} \left\{ I(l=l'\neq k')\mathbb{E}(x_{l,i}^2 x_{k',j}^2) + I(l=l'=k')\mathbb{E}(x_{k',i}^2 x_{k',j}^2) \right\} \\ &= n\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2, \end{aligned} \tag{A.10}$$

$$n^2\mathbb{E}(\bar{x}_j^2 x_{k,i}^2) = n\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2. \tag{A.11}$$

Substituting (A.5) and (A.9)–(A.11) into (A.8) gives

$$\mathbb{E}(\tilde{\sigma}_{ii}^s \tilde{\sigma}_{jj}^s) = \frac{n+1}{n-1}\sigma_{ii}\sigma_{jj} + \frac{2(n+2)}{n(n-1)}\sigma_{ij}^2. \tag{A.12}$$

Combining (A.7) and (A.12) gives (2.4). □

References

Bickel, P., Levina, E., 2008a. Covariance regularization by thresholding. Ann. Statist. 36, 2577–2604.
Bickel, P., Levina, E., 2008b. Regularized estimation of large covariance matrices. Ann. Statist. 36, 199–227.
Breiman, L., 1996. Heuristics of instability and stabilization in model selection. Ann. Statist. 24, 2350–2383.
Cai, T., Zhang, C.-H., Zhou, H., 2010. Optimal rates of convergence for covariance matrix estimation. Ann. Statist. 38, 2118–2144.
Cai, T., Zhou, H., 2010. Minimax estimation of large covariance matrices under ℓ_1-norm. Technical Report.
Donoho, D., Johnstone, I., 1995. Adapting to unknown smoothness via wavelet shrinkage. J. Amer. Statist. Assoc. 90, 1200–1224.
Efron, B., 1986. How biased is the apparent error rate of a prediction rule. J. Amer. Statist. Assoc. 81, 461–470.
Efron, B., 2004. The estimation of prediction error: covariance penalties and cross-validation. J. Amer. Statist. Assoc. 99, 619–632.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression (with discussion). Ann. Statist. 32, 407–499.
El Karoui, N., 2008. Operator norm consistent estimation of large dimensional sparse covariance matrices. Ann. Statist. 36, 2717–2756.
Frank, A., Asuncion, A., 2010. UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml.
Furrer, R., Bengtsson, T., 2007. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. J. Multivariate Anal. 98, 227–255.
Huang, J., Liu, N., Pourahmadi, M., Liu, L., 2006. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika 93, 85–98.
Johnstone, I., 2001. On the distribution of the largest eigenvalue in principal components analysis. Ann. Statist. 29, 295–327.
Lam, C., Fan, J., 2007. Sparsistency and rates of convergence in large covariance matrix estimation. Ann. Statist. 37, 4254–4278.
Rothman, A., Levina, E., Zhu, J., 2009. Generalized thresholding of large covariance matrices. J. Amer. Statist. Assoc. 104, 177–186.
Rothman, A., Levina, E., Zhu, J., 2010. A new approach to Cholesky-based covariance regularization in high dimensions. Biometrika 97, 539–550.
Shen, X., Ye, J., 2002. Adaptive model selection. J. Amer. Statist. Assoc. 97, 210–221.
Stein, C., 1981. Estimation of the mean of a multivariate normal distribution. Ann. Statist. 9 (6), 1135–1151.
Wu, W., Pourahmadi, M., 2009. Banding sample autocovariance matrices of stationary processes. Statist. Sinica 19, 1755–1768.
Zou, H., Hastie, T., Tibshirani, R., 2007. Orn the degrees of freedom of the lasso. Ann. Statist. 35, 2173–2192.