COVARIANCE REGULARIZATION BY THRESHOLDING

By Peter J. Bickel * and Elizaveta Levina †

University of California, Berkeley and University of Michigan

This paper considers regularizing a covariance matrix of p variables estimated from n observations, by hard thresholding. We show that the thresholded estimate is consistent in the operator norm as long as the true covariance matrix is sparse in a suitable sense, the variables are Gaussian or sub-Gaussian, and $(\log p)/n \to 0$, and obtain explicit rates. The results are uniform over families of covariance matrices which satisfy a fairly natural notion of sparsity. We discuss an intuitive resampling scheme for threshold selection and prove a general cross-validation result that justifies this approach. We also compare thresholding to other covariance estimators in simulations and on an example from climate data.

1. Introduction. Estimation of covariance matrices is important in a number of areas of statistical analysis, including dimension reduction by principal component analysis (PCA); classification by linear or quadratic discriminant analysis (LDA and QDA); establishing independence and conditional independence relations in the context of graphical models; and setting confidence intervals on linear functions of the means of the components. In recent years, many application area where these tools are used have been dealing with very high-dimensional datasets, and sample sizes can be very small relative to dimension. Examples include genetic data, brain imaging, spectroscopic imaging, climate data, and many others.

It is well known by now that the empirical covariance matrix for samples of size n from a p-variate Gaussian distribution, $\mathcal{N}_p(\mu, \Sigma_p)$, is not a good estimator of the population covariance if p is large. Many results in random matrix theory illustrate this, from the classical Marĉenko-Pastur law [24] to the more recent work of Johnstone and his students on the theory of the largest eigenvalues [12, 20, 25] and associated eigenvectors [21]. However, with the exception of a method for estimating the covariance spectrum [11], these probabilistic results do not offer alternatives to the sample covariance matrix.

*Supported by a grant from the NSF (DMS-0605236).

†Supported by grants from the NSF (DMS-0505424) and the NSA (MSPF-04Y-120).

AMS 2000 subject classifications: Primary 62H12; secondary 62F12, 62G09

Keywords and phrases: covariance estimation, regularization, sparsity, thresholding, large p small n, high dimension low sample size

Alternative estimators for large covariance matrices have therefore attracted a lot of attention recently. Two broad classes of covariance estimators have emerged: those that rely on a natural ordering among variables, and assume that variables far apart in the ordering are only weakly correlated, and those invariant to variable permutations. The first class includes regularizing the covariance matrix by banding or tapering [2, 3, 14], which we will discuss below. It also includes estimators based on regularizing the Cholesky factor of the inverse covariance matrix. These methods use the fact that the entries of the Cholesky factor have a regression interpretation, which allows application of regression regularization tools such as the lasso and ridge penalties [18], or the nested lasso penalty [23] specifically designed for the ordered variables situation. Banding the Cholesky factor has also been proposed [3, 29]. These estimators are appropriate for a number of applications with ordered data (time series, spectroscopy, climate data). For climate applications and other spatial data, since there is no total ordering on the plane, applying the Cholesky factor methodology is problematic, but as long as there is a metric on variable indexes (in this case, geographical distance), banding or tapering the covariance matrix can be applied.

However, there are many applications, e.g., gene expression arrays, where there is no notion of distance between variables at all. These applications require estimators invariant under variable permutations. Shrinkage estimators are in this category and have been proposed early on [7, 17]. More recently, Ledoit and Wolf [22] proposed an estimator where the optimal amount of shrinkage is estimated from data. Shrinkage estimators shrink the over-dispersed sample covariance eigenvalues, but they do not change the eigenvectors, which are also inconsistent [21], and do not result in sparse estimators. Several recent papers [5, 26, 30], construct a sparse permutation-invariant estimate of the *inverse* of the covariance matrix, also known as the concentration or precision matrix. Sparse concentration matrices are of interest in graphical models, since zero partial correlations imply a graph structure. The common approach of [5, 26, 30] is to add an L_1 (lasso) penalty on the entries of the concentration matrix to the normal likelihood, which results in shrinking some of the elements of the inverse to zero. In [26], it was shown that this method has a rate of convergence that is driven by $(\log p)/n$ and the sparsity of the truth. Computing this estimator is non-trivial for high dimensions and can be achieved either via a semi-definite programming algorithm [5, 30] or by using the Cholesky decomposition to re-parametrize the concentration matrix [26], but all of these are very computationally intensive. In specific applications, there have been other permutation-invariant approaches that use different notions of sparsity: Zou

et al. [31] apply the lasso penalty to loadings in PCA to achieve sparse representation; d'Aspremont et al. [6] compute sparse principal components by semi-definite programming; Johnstone and Lu [21] regularize PCA by moving to a sparse basis and thresholding; and Fan et al. [13] impose sparsity on the covariance via a factor model, which is often appropriate in finance applications.

In this paper, we propose thresholding of the sample covariance matrix as a simple and permutation-invariant method of covariance regularization. This idea has been simultaneously and independently developed by El Karoui [10], who studied it under a special notion of sparsity called β-sparsity (see details in Section 2.4). Here we develop a natural permutation-invariant notion of sparsity which, though more specialized than El Karoui's, seems easier to analyze and parallels the treatment in [3] which defines a class of models where banding is appropriate. Bickel and Levina [3] showed that, uniformly over the class of approximately "bandable" matrices, the banded estimator is consistent in the operator norm (also known as the matrix 2-norm, or spectral norm) for Gaussian data as long as $(\log p)/n \to 0$.

Here we show consistency of the thresholded estimator in the operator norm, uniformly over the class of matrices that satisfy our notion of sparsity, as long as $(\log p)/n \to 0$, and obtain explicit rates of convergence. There are various arguments to show that convergence in the operator norm implies convergence of eigenvalues and eigenvectors [3, 10], so this norm is particularly appropriate for PCA applications. The rate we obtain is slightly worse than the rate of banding when the variables are ordered, but the difference is not sharp. This is expected, since in the situation when variables are ordered, banding takes advantage of the underlying true structure. Thresholding, on the other hand, is applicable to many more situations. In fact, our treatment is in many respects similar to the pioneering work on thresholding of Donoho and Johnstone [8] and the recent work of Johnstone and Silverman [19] and Abramovich et al. [1].

The rest of this paper is organized as follows. In Section 2 we introduce the thresholding estimator and our notion of sparsity, prove the convergence result, and compare to results of El Karoui (Section 2.4) and to banding (Section 2.5). In Section 3, we discuss a cross-validation approach to threshold selection, which is novel in this context, and prove a cross-validation result of general interest. Section 4 gives simulations comparing several permutation-invariant estimators and banding. Section 5 gives an example of thresholding estimator applied to climate data, and Section 6 concludes with discussion.

2. Asymptotic results for thresholding. We start by setting up notation. We write $\lambda_{\max}(M) = \lambda_1(M) \geq \cdots \geq \lambda_p(M) = \lambda_{\min}(M)$ for the eigenvalues of a matrix M. Following the notation of [3], we define, for any $0 \leq r, s \leq \infty$ and a $p \times p$ matrix M,

$$\|M\|_{(r,s)} \equiv \sup \{\|M\boldsymbol{x}\|_s : \|\boldsymbol{x}\|_r = 1\}, \tag{1}$$

where $\|\boldsymbol{x}\|_r^r = \sum_{j=1}^p |x_j|^r$. In particular, we write $\|M\| = \|M\|_{(2,2)}$ for the operator norm, which for a symmetric matrix is given by

$$\|M\| = |\lambda_{\max}(M)|.$$

For symmetric matrices, we have (see e.g., [15]),

$$\|M\| \leq (\|M\|_{(1,1)}\|M\|_{(\infty,\infty)})^{1/2} = \|M\|_{(1,1)} = \max_j \sum_i |m_{ij}| \ . \tag{2}$$

We also use the Frobenius matrix norm,

$$\|M\|_F^2 = \sum_{i,j} m_{ij}^2 = \mathrm{tr}(MM^T).$$

We define the thresholding operator by

$$T_s(M) = [m_{ij}\ 1(|m_{ij}| \geq s)], \tag{3}$$

which we refer to as *M thresholded at s*. Note that T_s preserves symmetry and is invariant under permutations of variable labels, but does not necessarily preserve positive definiteness. However, if

$$\|T_s - T_0\| \leq \varepsilon \text{ and } \lambda_{\min}(M) > \varepsilon, \tag{4}$$

then $T_s(M)$ is necessarily positive definite, since for all vectors $\boldsymbol{v}$ with $\|v\|_2 = 1$ we have $\boldsymbol{v}^T T_s M \boldsymbol{v} \geq \boldsymbol{v}^T M \boldsymbol{v} - \varepsilon \geq \lambda_{\min}(M) - \varepsilon > 0$.

2.1. *A uniformity class of covariance matrices.* Recall that the banding operator was defined in [3] as $B_k(M) = [m_{ij}\ 1(|i-j| \leq k)]$. The uniformity class of "approximately bandable" covariance matrices is defined by

$$\begin{aligned}\mathcal{U}(\varepsilon_0, \alpha, C) = \big\{\Sigma : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Ck^{-\alpha} \text{ for all } k > 0, \\ \text{and } 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0\big\}.\end{aligned} \tag{5}$$

Here we define the uniformity class of covariance matrices invariant under permutations by

$$\mathcal{U}_\tau\big(q, c_0(p), M\big) = \{\Sigma : \sigma_{ii} \le M, \ \sum_{j=1}^{p} |\sigma_{ij}|^q \le c_0(p), \text{ for all } i\} \ ,$$

for $0 \le q < 1$. Thus, if $q = 0$,

$$\mathcal{U}_\tau\big(0, c_0(p), M\big) = \{\Sigma : \sigma_{ii} \le M, \ \sum_{j=1}^{p} 1(\sigma_{ij} \neq 0) \le c_0(p)\} \ ,$$

a class of sparse matrices. We will mainly write c_0 for $c_0(p)$ in the future. Note that

$$\lambda_{\max}(\Sigma) \le \max_i \sum_j |\sigma_{ij}| \le M^{1-q} c_0(p) \ ,$$

by the bound (2). Thus, if we define,

$$\mathcal{U}_\tau\big(q, c_0(p), M, \varepsilon_0\big) = \{\Sigma : \Sigma \in \mathcal{U}_\tau\big(q, c_0(p), M\big) \text{ and } \lambda_{\min}(\Sigma) \ge \varepsilon_0 > 0\} \ ,$$

we have a class analogous to (5).

Naturally, there is a class of covariance matrices that satisfies both banding and thresholding conditions. Define a subclass of $\mathcal{U}(\varepsilon_0, \alpha, C)$ by,

$$\begin{aligned}\mathcal{V}(\varepsilon_0, \alpha, C) = \big\{\Sigma : &|\sigma_{ij}| \le C|i-j|^{-(\alpha+1)}, \text{ for all } i, \ j: \ |i-j| \ge 1, \\ &\text{and } 0 < \varepsilon_0 \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le 1/\varepsilon_0\big\}.\end{aligned}$$

for $\alpha > 0$. Evidently,

$$\mathcal{V}(\varepsilon, \alpha, C) \subset \mathcal{U}(\varepsilon_0, \alpha, C_1)$$

for $C_1 \le C(1 + 1/\alpha)$.

On the other hand, $\Sigma \in \mathcal{V}(\varepsilon_0, \alpha, C)$ implies

$$\sum_j |\sigma_{ij}|^q \le \varepsilon_0^{-q} + C\frac{(\alpha+1)q}{(\alpha+1)q - 1} \ ,$$

so that for a suitable choice of c_0 and M,

$$\mathcal{V}(\varepsilon_0, \alpha, C) \subset \mathcal{U}_\tau(q, c_0, M)$$

for $q > \frac{1}{\alpha+1}$.

2.2. *Main result.* Suppose we observe n i.i.d. p-dimensional observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ distributed according to a distribution F, with $E\boldsymbol{X} = 0$ (without loss of generality), and $E(\boldsymbol{X}\boldsymbol{X}^T) = \Sigma$. We define the empirical (sample) covariance matrix by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \big(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\big)\big(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\big)^T, \tag{6}$$

where $\bar{\boldsymbol{X}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i$, and write $\hat{\Sigma} = [\hat{\sigma}_{ij}]$.

We have the following result which parallels the banding result (Theorem 1) of Bickel and Levina [3].

THEOREM 1. *Suppose F is Gaussian. Then, uniformly on $\mathcal{U}_\tau\big(q, c_0(p), M\big)$, for sufficiently large M', if*

$$t_n = M'\sqrt{\frac{\log p}{n}}\ , \tag{7}$$

and $\frac{\log p}{n} = o(1)$, then

$$\|T_{t_n}(\hat{\Sigma}) - \Sigma\| = O_P\left(c_0(p)\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\right)\ ,$$

and uniformly on $\mathcal{U}_\tau\big(q, c_0(p), M, \varepsilon_0\big)$,

$$\|(T_{t_n}(\hat{\Sigma}))^{-1} - \Sigma^{-1}\| = O_P\left(c_0(p)\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\right)\ .$$

Proof of Theorem 1: Recall that, without loss of generality, we assumed $E\boldsymbol{X} = \boldsymbol{0}$. Begin with the decomposition,

$$\hat{\Sigma} = \hat{\Sigma}^0 - \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^T\ , \tag{8}$$

where

$$\hat{\Sigma}^0 \equiv [\hat{\sigma}^0_{ij}] = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T\ .$$

Note that, by (8)

$$\max_{i,j}|\hat{\sigma}_{ij} - \sigma_{ij}| \le \max_{i,j}|\hat{\sigma}^0_{ij} - \sigma_{ij}| + \max_{i,j}|\bar{X}_i\bar{X}_j|\ . \tag{9}$$

The second term above is, by the union sum inequality,

$$\max_i |\bar{X}_i|^2 = O_P\left(\frac{\log p}{n}\right) , \tag{10}$$

since F is Gaussian and $\sigma_{ii} \leq M$ for all i. By a result of Saulis and Statulevičius [27] adapted for this context in Lemma 3 of [3], and $\sigma_{ii} \leq M$ for all i,

$$P\big[\max_{i,j} |\hat{\sigma}^0_{ij} - \sigma_{ij}| \geq t\big] \leq p^2 e^{-\delta n t^2} , \tag{11}$$

if $t = o(1)$.

We now recap an argument of Donoho and Johnstone [8]. Bound,

$$\|T_t(\hat{\Sigma}^0) - \Sigma\| \leq \|T_t(\Sigma) - \Sigma\| + \|T_t(\hat{\Sigma}^0) - T_t(\Sigma)\| .$$

The first term above is bounded by

$$\max_i \sum_{j=1}^p |\sigma_{ij}|\ 1(|\sigma_{ij}| \leq t) \leq t^{1-q} c_0(p) . \tag{12}$$

On the other hand,

$$\begin{aligned}
\|T_t(\hat{\Sigma}^0) - T_t(\Sigma)\| &\leq \max_i \sum_{j=1}^p |\hat{\sigma}^0_{ij}| 1(|\hat{\sigma}^0_{ij}| \geq t,\ |\sigma_{ij}| < t) \\
&\quad + \max_i \sum_{j=1}^p |\sigma_{ij}|\ 1(|\hat{\sigma}^0_{ij}| < t,\ |\sigma_{ij}| \geq t) \\
&\quad + \max_i \sum_{j=1}^p |\hat{\sigma}^0_{ij} - \sigma_{ij}|\ 1(|\hat{\sigma}^0_{ij}| \geq t,\ |\sigma_{ij}| \geq t) = \mathrm{I} + \mathrm{II} + \mathrm{III} .
\end{aligned} \tag{13}$$

Using (11), we have

$$\mathrm{III} \leq \max_{i,j} |\hat{\sigma}^0_{ij} - \sigma_{ij}|\ \max_i \sum_{j=1}^p |\sigma_{ij}|^q t^{-q} = O_P\left(c_0(p) t^{-q} \sqrt{\frac{\log p}{n}}\right) .$$

To bound term I, write

$$\begin{aligned}
\mathrm{I} &\leq \max_i \sum_{j=1}^p |\hat{\sigma}^0_{ij} - \sigma_{ij}|\ 1(|\hat{\sigma}^0_{ij}| \geq t,\ |\sigma_{ij}| < t) \\
&+ \max_i \sum_{j=1}^p |\sigma_{ij}|\ 1(|\sigma_{ij}| < t) \leq \mathrm{IV} + \mathrm{V} .
\end{aligned} \tag{14}$$

By (12),

$$\mathrm{V} \le t^{1-q} c_0(p) \ . \tag{15}$$

Now take $\gamma_1 \in (0,1)$. Then,

$$\begin{aligned}
\mathrm{IV} &\le \max_i \sum_{j=1}^{p} |\hat\sigma^0_{ij} - \sigma_{ij}|\ 1(|\hat\sigma^0_{ij}| \ge t,\ |\sigma_{ij}| \le \gamma_1 t) \\
&\quad + \max_i \sum_{j=1}^{p} |\hat\sigma^0_{ij} - \sigma_{ij}|\ 1(|\hat\sigma^0_{ij}| > t,\ \gamma_1 t < |\sigma_{ij}| < t,) \\
&\le \max_{i,j} |\hat\sigma^0_{ij} - \sigma_{ij}|\ \max_i N_i(1-\gamma_1) + c_0(p)(\gamma_1 t)^{-q} \max_{i,j} |\hat\sigma^0_{ij} - \sigma_{ij}| \ , \qquad (16)
\end{aligned}$$

where $N_i(a) \equiv \sum_{j=1}^{p}\ 1\big(|\hat\sigma^0_{ij} - \sigma_{ij}| > at\big)$. Note that, for some $\delta > 0$.

$$P[N_i(1-\gamma_1) > 0] = P[\max_{i,j} |\hat\sigma^0_{ij} - \sigma_{ij}| > (1-\gamma_1)t] \le p^2 e^{-n\delta(1-\gamma_1)^2 t^2} \ , \tag{17}$$

if $t = o(1)$, uniformly on $\mathcal{U}$. By (17) and (15), and $0 < \gamma_1 < 1$, if

$$2\log p - n\delta t^2 \to -\infty \ , \tag{18}$$

then,

$$\mathrm{IV} = O_P\big(c_0(p) t^{-q} \sqrt{\frac{\log p}{n}}\big) \ , \tag{19}$$

and, by (9) and (12),

$$\mathrm{I} = O_P\left(c_0(p) t^{-q} \sqrt{\frac{\log p}{n}} + c_0(p) t^{1-q}\right) \ . \tag{20}$$

Similarly, taking $\gamma_2 > 1$, we have

$$\begin{aligned}
\mathrm{II} &\le \max_i \sum_{j=1}^{p} |\sigma_{ij}|\ 1(|\hat\sigma^0_{ij}| < t,\ t \le |\sigma_{ij}| \le \gamma_2 t) \\
&\quad + \max_i \sum_{j=1}^{p} |\sigma_{ij}|\ 1(|\hat\sigma^0_{ij} - \sigma_{ij}| \ge \gamma_2 t) \\
&\le (\gamma_2 t)^{1-q} c_0(p) + M \max_i N_i(\gamma_2) = O_P(t^{1-q} c_0(p)) \ . \qquad (21)
\end{aligned}$$

Combining (20) and (21) and choosing t as in (7) establishes the first claim of the theorem. The second claim follows since

$$\big\| \big[T_{t_n}\big(\hat\Sigma\big)\big]^{-1} - \Sigma^{-1} \big\| = \Omega_P\big(\|T_{t_n}\big(\hat\Sigma\big) - \Sigma\|\big)$$

uniformly on $\mathcal{U}_\tau(q, c_0(p), M, \varepsilon_0)$, where $A = \Omega_P(B)$ means $A = O_P(B)$ and $B = O_P(A)$. □

THEOREM 2. *Suppose F is Gaussian. Then, uniformly on $\mathcal{U}_\tau(q, c_0(p), M)$, if $t = M'\sqrt{\frac{\log p}{n}}$ and M' is sufficiently large,*

$$\frac{1}{p}\|T_t(\hat{\Sigma}) - \Sigma\|_F^2 = O_P\Big(c_0(p)\frac{\log p}{n}\Big)^{1-q/2} . \tag{22}$$

An analogous result holds for the inverse on $\mathcal{U}_\tau(q, c_0(p), M, \varepsilon_0)$.

Proof of Theorem 2. The proof is essentially the same as for Theorem 1. We need to bound,

$$\sum_{a,b} \big(\hat{\sigma}_{ab}1(|\hat{\sigma}_{ab}| \geq t) - \sigma_{ab}\big)^2 .$$

As before,

$$\sum_{a,b} \sigma_{ab}^2 1(|\sigma_{ab}| < t) \leq t^{2-q} p c_0(p) . \tag{23}$$

Similarly, for instance,

$$\sum_{a,b} \big(\hat{\sigma}_{ab} - \sigma_{ab}\big)^2 1\big(|\hat{\sigma}_{ab}| \geq t, |\sigma_{ab}| \geq t\big) \leq t^{-q} p c_0(p) \frac{\log p}{n}\big(1 + o_P(1)\big) . \tag{24}$$

The theorem follows by from putting (23), (24) and the other remainder terms together. It is clear from the argument that, by restricting the result from the class $\mathcal{U}_\tau(q, c_0(p), M)$ to properly chosen Σ's, we can change O_P into Ω_P. □.

2.3. *The non-Gaussian case.* We consider two cases here. If, for some $\eta > 0$,

$$Ee^{tX_{ij}^2} \leq K < \infty \text{ for all } |t| \leq \eta_j, \text{ for all } i, j$$

then the proof goes through verbatim, since result (11) still holds. The bound on $\max_i |\bar{X}_i|^2$ will always be at least the squared rate of $\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}|$, hence we do not need normality for (10).

In the second case, if we have, for some $\gamma > 0$,

$$E|X_{ij}|^{2(1+\gamma)} \leq K \text{ for all } i, j,$$

then by Markov's inequality

$$P[|\hat{\sigma}_{ij} - \sigma_{ij}| \geq t] \leq KC(\gamma)\frac{n^{-(1+\gamma)/2}}{t^{1+\gamma}} . \tag{25}$$

Thus the bound (11) becomes

$$p^2 KC(\gamma)\frac{n^{-(1+\gamma)/2}}{t^{1+\gamma}}$$

and hence,

$$\max_{i,j}|\hat{\sigma}^0_{ij}-\sigma_{ij}| = O_P\left(\frac{p^{2/(1+\gamma)}}{n^{1/2}}\right) .$$

Therefore, taking $t_n = M\frac{p^{2/(1+\gamma)}}{n^{1/2}}$, we find that,

$$\|T_{t_n}(\hat{\Sigma})-\Sigma\| = O_P\left(c_0(p)\left(\frac{p^{2/(1+\gamma)}}{n^{1/2}}\right)^{1-q}\right) . \tag{26}$$

which is we expect minimax though this needs to be checked.

2.4. *Comparison to thresholding results of El Karoui.* El Karoui [10] shows as a special case that if,

(i) $E|X_j|^r < \infty$ for all r, $1 \le j \le p$.
(ii) $\sigma_{jj} \le M < \infty$ for all j.
(iii) If $\sigma_{ij} \neq 0$, $|\sigma_{ij}| > Cn^{-\alpha_0}$, $0 < \alpha_0 < \frac{1}{2} - \delta_0 < \frac{1}{2}$.
(iv) Σ is β-sparse, $\beta = \frac{1}{2} - \eta$, $\eta > 0$.
(v) $\frac{p}{n} \to c \in (0, \infty)$.

Then, if $t_n = Cn^{-\alpha}$, $\alpha = \frac{1}{2} - \delta_0 > \alpha_0$

$$\|T_{t_n}(\hat{\Sigma})-\Sigma\| \xrightarrow{\text{a.s.}} 0 . \tag{27}$$

El Karoui's notion of β-sparsity is such that our case $q = 0$ is β-sparse with $c_0(p) = Kp^\beta$. Our results yield a rate of

$$O_P\left(\frac{p^{\beta+2/(1+\gamma)}}{n^{1/2}}\right) ,$$

for γ arbitrarily large. Since $\beta < \frac{1}{2}$ by assumption and $p \asymp n$ we see that our result implies (27) under (i), (ii), (iv), (v) and our notion of sparsity. Thus, our result is stronger than his in the all moments case, again under our stronger notion of sparsity. El Karoui's full result, in fact, couples a maximal value of r in (i) with the largest possible value of β. Unfortunately, this coupling involves (iii) which we do not require. Nevertheless, his result implies the corresponding consistency results of ours, if (iii) is ignored, when only existence of a finite set of moments is assumed. However, according to El Karoui (personal communication), (iii) is not needed for (27) in the case when our sparsity condition holds.

2.5. *Comparison to banding results of Bickel and Levina.* Comparison is readily possible on $\mathcal{V}(\varepsilon_0, \alpha, C)$. By Theorem 1 of [3] the best rate achievable using banding is

$$O_P\left(\left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(\alpha+1)}}\right) .$$

On the other hand, by our Theorem 1, thresholding yields,

$$O_P\left(\left(\frac{\log p}{n}\right)^{\frac{1-q}{2}}\right) ,$$

for $q > \frac{1}{\alpha+1}$. Comparing exponents, we see that, banding is slightly better in the situation where labels are meaningful, since we must have,

$$1 - q < \frac{\alpha}{\alpha+1} .$$

However, since $1-q$ can be arbitrarily close to $\frac{\alpha}{\alpha+1}$ the difference is not sharp. Not surprisingly, as $\alpha \to \infty$, the genuinely sparse case, the bounds both approach $\left(\frac{\log p}{n}\right)^{1/2}$.

3. Choice of threshold. The question of threshold selection seems to be hard to answer analytically. In fact, the $\hat{\sigma}_{ij}$ have variances which depend on the distribution of (X_i, X_j) through higher order moments so it may in fact make sense to threshold differentially. We conjecture that this would not make much difference if we assume second and fourth moments bounded above and below. Ignoring this issue, we propose a cross-validation method analogous to the one used by Bickel and Levina [3] but made using the Frobenius metric which enables us to successfully analyze it.

3.1. *Method.* Split the sample randomly into two pieces of size n_1 and n_2 where a choice to be "justified" theoretically is $n_1 = n\left(1 - \frac{1}{\log n}\right)$, $n_2 = \frac{n}{\log n}$ and repeat this N times. Let $\hat{\Sigma}_1^{(\nu)}$, $\hat{\Sigma}_2^{(\nu)}$ be the empirical covariance matrices based on the n_1 and n_2 observations respectively from the ν-th split. Form

$$\hat{R}(s) = \frac{1}{N}\sum_{\nu=1}^{N} \|T_s(\hat{\Sigma}_1^{(\nu)}) - \hat{\Sigma}_2^{(\nu)}\|_F^2 , \tag{28}$$

and choose $\hat{s}$ to minimize $\hat{R}(s)$ (in practice for $s \geq \varepsilon_n \to 0$, $\varepsilon_n \asymp \frac{\log p}{n}$). We will show that, under the conditions of Theorem 2, and an additional condition, we have for $q \geq 0$

$$\frac{1}{p}\|T_{\hat{s}}(\hat{\Sigma}) - \Sigma\|_F^2 = O_P\left[\left(\frac{\log p}{n}\right)^{1-q/2} c_0(p)\right] . \tag{29}$$

uniformly on the appropriate uniformity class. Claim (28) is weaker than the desired

$$\|T_{\hat{s}}(\hat{\Sigma}) - \Sigma\|_{(2,2)} = O_P\left[\left(\frac{\log p}{n}\right)^{1-q} c_0(p)\right] , \tag{30}$$

but we would expect the normalized Frobenius norm statement (29) to be a reasonable proxy for (30).

We introduce the additional condition in terms of a uniformity class for Gaussian mean $\mathbf{0}$ p-dimensional distributions of $\boldsymbol{X}_1$. Let $\Delta = [X_{1a}X_{1b} - \sigma_{ab}]_{p\times p}$, and define

$$\mathcal{W} = \left\{\Sigma \in U_\tau(q, c_0(p), M) : E \max_{1\le j\le J} \|V_j * \Delta\|_F^2 \le C\rho(J)\right\} ,$$

for all $V_1, \ldots, V_J$ $p\times p$ symmetric, $\|V_j\|_F = 1$. Here $*$ denotes Schur (elementwise) matrix product and C and $\rho(\cdot)$ are independent of Σ. For convenience, we will use C as a generic constant throughout.

Note that under exponential mixing conditions on $\{X_{1a}X_b : 1 \le a \le p,\ 1 \le b \le p\}$ we may be able to estimate $\rho(J)$ using the results of [27] Section 4.4. Typically we can expect $\rho(J) \sim J^\alpha$, $\alpha > 0$. For $\boldsymbol{X}_1 = (X_{11}, \ldots, X_{1p})$ a stationary Markov chain obeying exponential mixing we would even expect $\rho(J)$ to be slowly varying.

We begin with two essential technical results of possibly independent interest.

3.2. *An inequality.* We note an inequality derivable from a classic one of Pinelis – see [28] for instance.

PROPOSITION 1. *Let $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n$ be i.i.d. p-variate vectors with $E|\boldsymbol{U}_1|^2 \le K$, $E\boldsymbol{U}_1 = \mathbf{0}$. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_J$ be fixed p-variate vectors of length 1. Define for $\boldsymbol{x} \in R^p$*

$$\|\boldsymbol{x}\|_{\boldsymbol{v}} = \max_{1\le j\le J} |\boldsymbol{v}_j^T\boldsymbol{x}| .$$

Then,

$$E\|\sum_{i=1}^n \boldsymbol{U}_i\|_{\boldsymbol{v}}^2 \le C\, n\, \log J\, E\|\boldsymbol{U}_1\|_{\boldsymbol{v}}^2 , \tag{31}$$

where C is an absolute constant.

Proof of Proposition 1: By symmetrization,

$$E\|\sum_{i=1}^n \boldsymbol{U}_i\|_{\boldsymbol{v}}^2 \le 2E \max_j \left(\sum_{i=1}^n \varepsilon_i |(\boldsymbol{U}_i - \boldsymbol{U}_i')^T\boldsymbol{v}_j|\right)^2$$

where $\boldsymbol{U}_i'$ are i.i.d. as $\boldsymbol{U}_i$ and independent of $\boldsymbol{U}_i$, and $\{\varepsilon_i\}$ are ± 1 with probability $1/2$ and independent of $|(\boldsymbol{U}_i - \boldsymbol{U}_i')^T \boldsymbol{v}_j|$. Let

$$W_{ij} = |(\boldsymbol{U}_i - \boldsymbol{U}_i')^T \boldsymbol{v}_j|, \ a_{ij} = \frac{W_{ij}}{(\sum_{i=1}^n W_{ij}^2)^{1/2}} \ .$$

Then,

$$E \max_{1\le j\le J} \big(\sum_{i=1}^n \varepsilon_i W_{ij}\big)^2 \le E\Big\{E\Big[\max_{1\le j\le J}(\sum_{i=1}^n a_{ij}\varepsilon_i)^2 \mid \{W_{ij} : 1\le i\le n,\ 1\le j\le J\}\Big]\times \max_{1\le j\le J}\sum_{i=1}^n W_{ij}^2\Big\} \le Cn\log JE \max_{1\le j\le J}\sum_{i=1}^n W_{ij}^2 \ ,$$

by Pinelis' inequality [28]. Thus

$$\begin{aligned} E \max_{1\le j\le J} \big(\sum_{i=1}^n \varepsilon_i W_{ij}\big)^2 &\le Cn\log JE \max_{1\le j\le J}\big((\boldsymbol{U}_i - \boldsymbol{U}_i')^T\boldsymbol{v}_j\big)^2 \\ &\le 2Cn\log JE \max_{1\le j\le J}(\boldsymbol{U}_1^T\boldsymbol{v}_j)^2 \ . \qquad \square \end{aligned}$$

3.3. *A general result on V-fold cross-validation.* We will prove our result for $N = 1$ in (28). The nature of our argument in Theorem 3 is such that it is fairly easy to see that it applies to each term of the sum in (28) and thus holds not just for the "sample splitting" ($N = 1$) procedure, but also for the general 2-fold cross-validation procedure that is given by (28), and in fact for more general V-fold cross-validation procedures.

Let $\boldsymbol{W}_1, \dots, \boldsymbol{W}_{n+B}$ be i.i.d. p-variate vectors with distribution P, with $E_P \boldsymbol{W} \equiv \boldsymbol{\mu}(P)$. Let $\hat{\boldsymbol{\mu}}_j$, $1 \le j \le J$ be estimates of $\boldsymbol{\mu}$ based on $\boldsymbol{W}_1, \dots, \boldsymbol{W}_n$. For convenience, in this section we write $|\boldsymbol{x}|^2 = \|\boldsymbol{x}\|_2^2 = \sum_{j=1}^p x_j^2$ and $(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{y}$. Let,

$$L(\boldsymbol{\mu}, \boldsymbol{d}) = |\boldsymbol{\mu} - \boldsymbol{d}|^2 \ .$$

The *oracle estimate* $\hat{\boldsymbol{\mu}}^o$ is defined by

$$\hat{\boldsymbol{\mu}}^o \equiv \arg\min_j |\boldsymbol{\mu}(P) - \hat{\boldsymbol{\mu}}_j|^2 \ .$$

The *sample splitting* estimate $\hat{\boldsymbol{\mu}}^c$ is defined as follows. Let

$$\bar{\boldsymbol{W}}_B = \frac{1}{B}\sum_{j=1}^B \boldsymbol{W}_{n+j} \ .$$

Then,

$$\hat{\boldsymbol{\mu}}^c \equiv \arg\min_j |\bar{\boldsymbol{W}}_B - \hat{\boldsymbol{\mu}}_j|^2 \ .$$

Here is our basic result which has in some form appeared in Gyorfi et al. [16] (Ch. 7, Theorem 7.1, p.101), Bickel et al. [4], and Dudoit and van der Laan [9]. The major public proof in [16] appears to be in error and does not directly apply to our case so we give the proof of our statement for completeness.

THEOREM 3. *Suppose,*

(A1) $|\hat{\boldsymbol{\mu}}^o - \boldsymbol{\mu}(P)|^2 = \Omega_p(r_n)$;
(A2) $E_P \max_{1\le j\le J} |(\boldsymbol{v}_j,\ \boldsymbol{W}_1 - \boldsymbol{\mu})|^2 \le C\rho(J)$ *for any set* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_J$ *of unit vectors in* R^p ;
(A3) $\rho(J_n)\frac{\log J_n}{B_n} = o(r_n)$.

Then,

$$|\hat{\boldsymbol{\mu}}^c - \boldsymbol{\mu}(P)|^2 = |\hat{\boldsymbol{\mu}}^o - \boldsymbol{\mu}(P)|^2(1 + o_P(1)) = \Omega_P(r_n) \ , \tag{32}$$

where $A = \Omega_P(B)$ *means that* $A = O_P(B)$ *and* $B = O_P(A)$.

As an example, note that if, for instance, $r_n = n^{-1+\delta}$, $\delta > 0$, $B = \frac{n}{\log n}$, $\rho(J) = J^\beta$, $J = n^\alpha$, $\alpha\beta < \delta$, our conditions are satisfied.

Proof of Theorem 3: By definition, writing $\boldsymbol{\mu} \equiv \boldsymbol{\mu}(P)$,

$$|\hat{\boldsymbol{\mu}}^c - \bar{\boldsymbol{W}}_B|^2 \le |\hat{\boldsymbol{\mu}}^o - \bar{\boldsymbol{W}}_B|^2 \ , \tag{33}$$

which is equivalent to,

$$2(\hat{\boldsymbol{\mu}}^c - \hat{\boldsymbol{\mu}}^0,\ \bar{\boldsymbol{W}}_B - \boldsymbol{\mu}) \ge |\hat{\boldsymbol{\mu}}^c - \boldsymbol{\mu}|^2 - |\hat{\boldsymbol{\mu}}^o - \boldsymbol{\mu}|^2 \ . \tag{34}$$

But,

$$|(\hat{\boldsymbol{\mu}}^c - \hat{\boldsymbol{\mu}}^o,\ \bar{\boldsymbol{W}}_B - \boldsymbol{\mu})| \le |(\hat{\boldsymbol{\mu}}^c - \boldsymbol{\mu},\ \bar{\boldsymbol{W}}_B - \boldsymbol{\mu})| + |(\hat{\boldsymbol{\mu}}^o - \boldsymbol{\mu},\ \bar{\boldsymbol{W}}_B - \boldsymbol{\mu})| \ , \tag{35}$$

Now, let

$$\hat{\boldsymbol{\nu}}_j = \frac{\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}}{|\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}|}.$$

Then we have

$$|(\hat{\boldsymbol{\mu}}^c - \boldsymbol{\mu},\ \bar{\boldsymbol{W}}_B - \boldsymbol{\mu})| \le |\hat{\boldsymbol{\mu}}^o - \boldsymbol{\mu}| \max_{1\le j\le J} \left|(\hat{\boldsymbol{\nu}}_j,\ \bar{\boldsymbol{W}}_B - \boldsymbol{\mu})\right| \ , \tag{36}$$

and similarly for the other term. Now, by Proposition 1 and assumption (A2),

$$E \max_{1\le j\le J} |(\hat{\boldsymbol{\nu}}_j,\ \bar{\boldsymbol{W}}_B - \boldsymbol{\mu})|^2 \le C\frac{\log J}{B}\rho(J) \tag{37}$$

where C is used generically. Therefore, after some algebra and Cauchy-Schwartz, by (33),

$$|\hat{\boldsymbol{\mu}}^c - \boldsymbol{\mu}|^2 \le O_P\Big(\frac{\log^{1/2} J}{B^{1/2}}\rho^{1/2}(J)\Big)\big(|\hat{\boldsymbol{\mu}}^c - \boldsymbol{\mu}| + |\hat{\boldsymbol{\mu}}^o - \boldsymbol{\mu}|\big) + |\hat{\boldsymbol{\mu}}^o - \boldsymbol{\mu}|^2\ . \tag{38}$$

Letting $|\hat{\boldsymbol{\mu}}^c - \boldsymbol{\mu}|^2 = a_n$ we can rewrite (34) as

$$a_n \le C\frac{\log^{1/2} J}{B^{1/2}}\rho^{1/2}(J)(a_n^{1/2} + r_n^{1/2}) + r_n\ , \tag{39}$$

with probability $1-\varepsilon(C)$, with $\varepsilon(C)\to 0$ as $C\to\infty$. Using (iii)

$$a_n \le a_n^{1/2} o_P(r_n^{1/2}) + r_n\big(1 + o_P(1)\big)\ .$$

But by definition,

$$a_n \ge r_n$$

and hence,

$$a_n^{1/2} \le o_P(r_n^{1/2}) + r_n^{1/2}\big(1 + o_P(1)\big)$$

and the theorem follows. □

Here is our main result establishing (29). As we indicated it is enough to consider $N = 1$, and for convenience write the observations as,

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m, \ldots, \boldsymbol{X}_{m+B}\ ,$$

where $n = m+B$. Form $\hat{\Sigma}^{(1)}$ the empirical covariance matrix of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and $\hat{\Sigma}^{(2)}$ that of $\boldsymbol{X}_{m+1}, \ldots, \boldsymbol{X}_{m+B}$ and the estimates, $T_o(\hat{\Sigma})$, $T_{\hat{t}}(\hat{\Sigma})$ corresponding to the oracle and statistician. By Theorem 2, it is clear that for suitable $\Sigma \in U_\tau$

$$\frac{1}{p}\|T_o(\hat{\Sigma}) - \Sigma\|_F^2 = \Omega_p\left(\left(\frac{\log p}{n}\right)^{1-q/2} c_0(p)\right)\ , \tag{40}$$

and $r_n = M'\sqrt{\frac{\log p}{n}}$ for appropriate M'. Let the optimizing $\hat{t}$ be, in fact, obtained by searching over a grid $\big\{j\sqrt{\frac{\log p}{n}}:\ 0 \le j \le J_n\big\}$. We see that as a consequence of Theorem 3 we can state,

THEOREM 4. *Suppose $\boldsymbol{X}_i$ are Gaussian, $\Sigma \in U_\tau(q, c_0(p), M)$, (40) holds, $J_n = n^\alpha$ and $q > 2\alpha$. Then*

$$\|T_{\hat{t}}(\hat{\Sigma}) - \Sigma\|_F = \|T_o(\hat{\Sigma}) - \Sigma\|_F\big(1 + o_P(1)\big) \ .$$

Even if (40) does not hold,

$$\frac{1}{p}\|T_{\hat{t}}(\hat{\Sigma}) - \Sigma\|_F^2 = O_P\left(\left(\frac{\log p}{n}\right)^{1-q/2}\right) \ .$$

The proof is immediate from Theorems 2 and 3.

4. Simulation results. The simulation results we present focus on comparing banding, thresholding, and two more permutation-invariant estimators: the sample covariance and the shrinkage estimator of Ledoit and Wolf [22]. We consider the AR(1) population covariance model,

$$\Sigma = [\sigma_{ij}] = [\rho^{|i-j|}] \tag{41}$$

with $\rho = 0.7$. The value of 0.7 was chosen so that the matrix is not very sparse (as would be the case with $\rho \leq 0.5$) but does have a fair number of very small entries (which would not be the case with ρ close to 1). For banding, we show results for the variables in their "correct" order, and permuted at random. All other estimators are invariant to variable permutations, so their results are the same for both of these scenarios. We consider three values of $p = 30, 100, 200$, and the sample size is fixed at $n = 100$.

Table 1 shows average losses and standard deviations over 100 replications, as measured by three different matrix norms (matrix 1-norm which we denote $\|\cdot\|_{(1,1)}$, operator, and Frobenius norms). We also report the absolute difference in the largest eigenvalue, $|\lambda_{\max}(\hat{\Sigma}) - \lambda_{\max}(\Sigma)|$, and the absolute value of the cosine of the angle between the estimated and true eigenvectors corresponding to the first eigenvalue. This assesses how accurate each of the estimators would be in estimating the first principal component.

The results in Table 1 show what one would expect: when banding is given the correct order of variables, it performs better than thresholding, since it is taking advantage of the underlying structure. When banding is given the variables in the wrong order, it performs poorly, often worse than the sample covariance matrix, and then thresholding is a much better choice. The Ledoit-Wolf estimator performs worse than thresholding by most measures, although it does well on estimating the largest eigenvalue. Note that the eigenvectors of the Ledoit-Wolf estimator are equal to the sample covariance eigenvectors.

TABLE 1
Averages and standard deviations over 100 replications of performance measures for AR(1) with $\rho = 0.7$.

p	Sample	Ledoit-Wolf	Banding	Banding Perm.	Thresholding
Matrix 1-norm					
30	3.87(0.70)	3.36(0.48)	2.54(0.47)	3.85(0.68)	3.28(0.50)
100	11.46(0.93)	7.99(0.47)	3.13(0.37)	5.05(0.07)	4.61(0.40)
200	22.00(1.36)	11.82(0.56)	3.34(0.34)	5.09(0.07)	4.99(0.06)
Operator norm					
30	1.95(0.43)	1.69(0.34)	1.38(0.28)	1.92(0.42)	1.90(0.36)
100	4.16(0.49)	3.06(0.21)	1.68(0.21)	4.63(0.03)	3.15(0.26)
200	6.68(0.64)	3.80(0.13)	1.80(0.18)	4.67(0.02)	3.64(0.18)
Frobenius norm					
30	3.19(0.37)	2.89(0.29)	2.42(0.26)	3.21(0.36)	3.42(0.32)
100	10.23(0.41)	8.16(0.22)	4.60(0.24)	13.80(0.01)	8.73(0.34)
200	20.24(0.5)	14.02(0.16)	6.61(0.28)	19.61(0.01)	13.79(0.27)
Abs. difference between true and estimated largest eigenvalue					
30	0.91(0.59)	0.46(0.42)	0.52(0.38)	0.84(0.58)	0.74(0.50)
100	2.86(0.60)	0.43(0.33)	0.38(0.26)	4.24(0.08)	1.07(0.45)
200	5.21(0.74)	0.42(0.29)	0.31(0.22)	4.23(0.07)	1.15(0.38)
Abs. cosine of the angle between true and estimated 1st PC					
30	0.77(0.25)	0.77(0.25)	0.81(0.12)	0.76(0.25)	0.70(0.15)
100	0.37(0.22)	0.37(0.22)	0.42(0.15)	0.10(0.04)	0.28(0.11)
200	0.27(0.18)	0.27(0.18)	0.26(0.12)	0.06(0.03)	0.18(0.09)

Table 2 shows the band width selected by the cross-validation procedure on correct and permuted orderings, and the threshold selection. Note that banding in permuted order always selects a diagonal model for both $p = 100$ and $p = 200$, and keeps almost all the entries at $p = 30$, both of which result in bad estimators. The selected threshold increases with dimension, which is expected since in higher dimensions one would need to regularize more. The selected band width also goes down with dimension (the decrease on our range of p's is not very large, but results over a wider range of dimensions in [3] show the same pattern more clearly).

TABLE 2
Averages and standard deviations over 100 replications of selected band width and threshold

p	Banding k	Banding Perm. k	Threshold t
30	4.36(0.70)	24.57(1.38)	0.33(0.04)
100	4.27(0.45)	0.00(0)	0.49(0.02)
200	4.22(0.42)	0.00(0)	0.55(0.01)

Figure 4 shows scree plots of the true eigenvalues and means, 2.5% and

97.5% percentiles of the estimates over 100 replications for $p = 100$. Interestingly, the results show that the sample covariance is very bad at estimating the leading eigenvalues, but better than thresholding on the middle part of the spectrum. The leading eigenvalues, however, are more important in applications like PCA. The Ledoit-Wolf's estimator does better on eigenvalues than on overall loss measures in Table 1. Banding in the correct order appears to do best on estimating the spectrum. For illustration purposes, scree plots from a single randomly selected realization are shown in Figure 4.

5. Climate data example. In this section, we illustrate the performance of the thresholded covariance estimator by applying it to climate data. The data are monthly mean temperatures recorded from January 1850 to June 2006; only the January data was used in the analysis below (157 observations). The region covered by the total of 2592 recording stations extends from -177.5 to 177.5 degrees longitude, and from -87.5 to 87.5 latitude. Not all the stations were in place for the entire period of 157 years; we do not impute the missing data in any way, but instead simply calculate spatial covariance from all the years available for any given pair of stations.

EOFs (empirical orthogonal functions) are frequently used in spatio-temporal statistics to represent patterns in spatial data. They are simply the principal components of the spatial covariance matrix, where observations over time are used as replications to calculate covariance between different spatial locations. EOFs are typically represented by spatial contour plots, which provide a visual illustration of which regions contribute the most to which principal components.

The plots in Figures 3 and 4 show the contour plots of the first four EOFs obtained, respectively, from the spatial sample covariance matrix (regular PCA), and from the thresholded spatial covariance matrix. We see that with thresholding, the first EOF essentially corresponds to Eurasia, and the second to North America, which the climate scientists agree should be separate. The regular PCA does not separate the continents.

6. Summary and Discussion. We have proposed and analyzed a regularization by thresholding approach to estimation of large covariance matrices. One of its biggest advantages is its simplicity – hard thresholding carries no computational burden, unlike many other methods for covariance regularization. A potential disadvantage is the loss of positive definiteness – but since we show that for a suitably sparse class of matrices the estimator is consistent as long as $(\log p)/n \to 0$, the estimator will be positive definite with probability tending to 1. We show consistency in the operator norm,

which guarantees consistency for principal components, hence we expect that PCA will be one of the most important applications of the method.

We have also provided theoretical justification for the cross-validation approach to selecting the threshold. While it was formulated in the context of hard thresholding, the general result is much more widely applicable; in particular, it applies to other covariance estimation methods that depend on selecting the tuning parameter, such as [3, 18, 23, 26] and others.

Acknowledgments. We thank Noureddine el Karoui (Statistics, UC Berkeley) for helpful discussions; Adam Rothman (Statistics, University of Michigan) for help with simulations; Donghui Yan (Statistics, UC Berkeley) and Serge Guillas (Mathematics, Georgia Tech University) for climate EOF plots; and Gabi Hegerl (Earth and Ocean Sciences, Duke University) for help in interpreting the climate plots.

REFERENCES

[1] Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics*, 34:584–653.

[2] Bickel, P. J. and 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(6):989–1010.

[3] Bickel, P. J. and Levina, E. (2006). Regularized estimation of large covariance matrices. *Ann. Statist.* To appear.

[4] Bickel, P. J., Ritov, Y., and Zakai, A. (2006). Some theory for generalized boosting algorithms. *J. Machine Learning Research*, 7:705–732.

[5] d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2007). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*. To appear.

[6] d'Aspremont, A., El Ghaoui, L., Jordan, M. I., and Lanckriet, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448.

[7] Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.*, 13(4):1581–1591.

[8] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

[9] Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154.

[10] El Karoui, N. (2007a). Operator norm consistent estimation of large dimensional sparse covariance matrices. Technical Report 734, UC Berkeley, Department of Statistics.

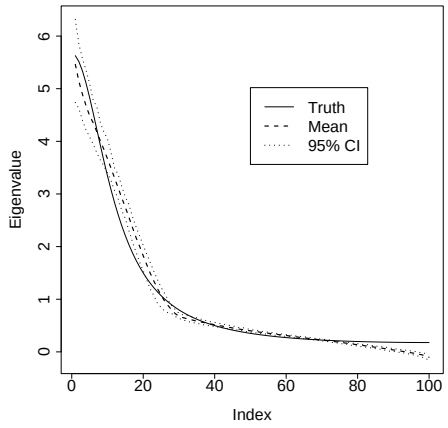
[11] El Karoui, N. (2007b). Spectrum estimation for large dimensional covariance matrices using random matrix theory. Technical report, UC Berkeley, Department of Statistics.

[12] El Karoui, N. (2007c). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Annals of Probability*, 35(2):663–714.

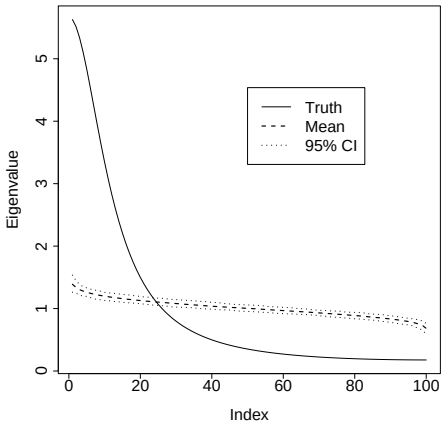
[13] Fan, J., Fan, Y., and Lv, J. (2006). High dimensional covariance matrix estimation using a factor model. Technical report, Princeton University. Manuscript.
[14] Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255.
[15] Golub, G. H. and Van Loan, C. F. (1989). *Matrix Computations*. The John Hopkins University Press, Baltimore, Maryland, 2nd edition.
[16] Gyorfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.
[17] Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.*, 8(3):586–597.
[18] Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.
[19] Johnstone, I. and Silverman, B. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics*, 33(4):1700–1752.
[20] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.*, 29(2):295–327.
[21] Johnstone, I. M. and Lu, A. Y. (2004). Sparse principal components analysis. *J. Amer. Statist. Assoc.* Tentatively accepted.
[22] Ledoit, O. and Wolf, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.
[23] Levina, E., Rothman, A. J., and Zhu, J. (2007). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Annals of Applied Statistics*. To appear.
[24] Marĉenko, V. A. and Pastur, L. A. (1967). Distributions of eigenvalues of some sets of random matrices. *Math. USSR-Sb*, 1:507–536.
[25] Paul, D. (2007). Asymptotics of the leading sample eigenvalues for a spiked covariance model. *Stat. Sinica*. To appear.
[26] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2007). Sparse permutation invariant covariance estimation. Technical Report 467, University of Michigan.
[27] Saulis, L. and Statulevičius, V. A. (1991). *Limit Theorems for Large Deviations*. Kluwer Academic Publishers, Dordrecht.
[28] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
[29] Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90:831–844.
[30] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.
[31] Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal components analysis. *Journal of Computational and Graphical Statistics*, 15:265–286.

Department of Statistics
University of California, Berkeley
Berkeley CA 94720-3860
E-mail: bickel@stat.berkeley.edu

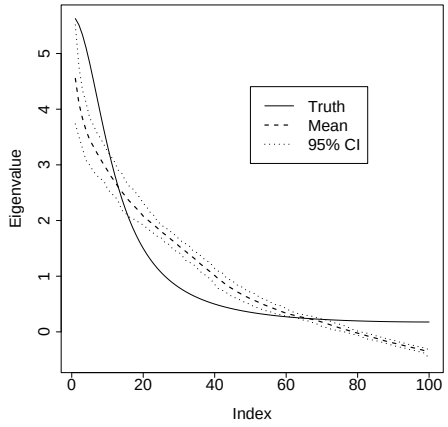
Department of Statistics
University of Michigan
Ann Arbor, MI 48109-1107
E-mail: elevina@umich.edu



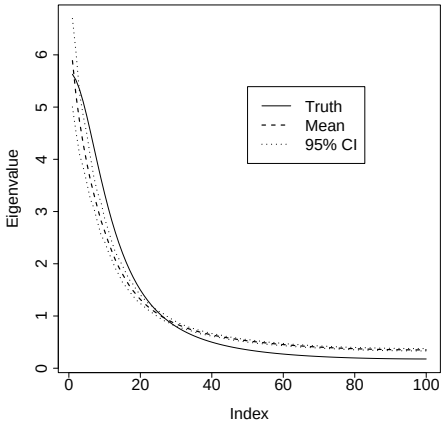
Banding



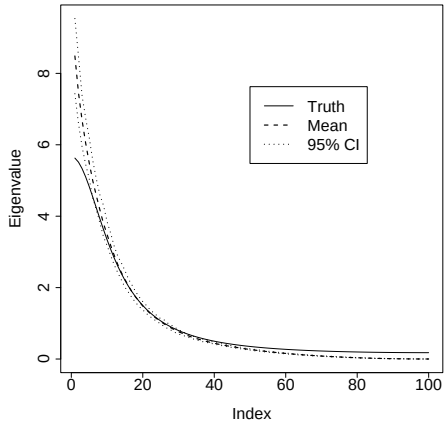
Banding Permuted



Thresholding



Ledoit-Wolf



Sample

FIG 1. *Scree plots: the mean estimated eigenvalues, their 2.5% and 97.5% percentiles, and the truth.*

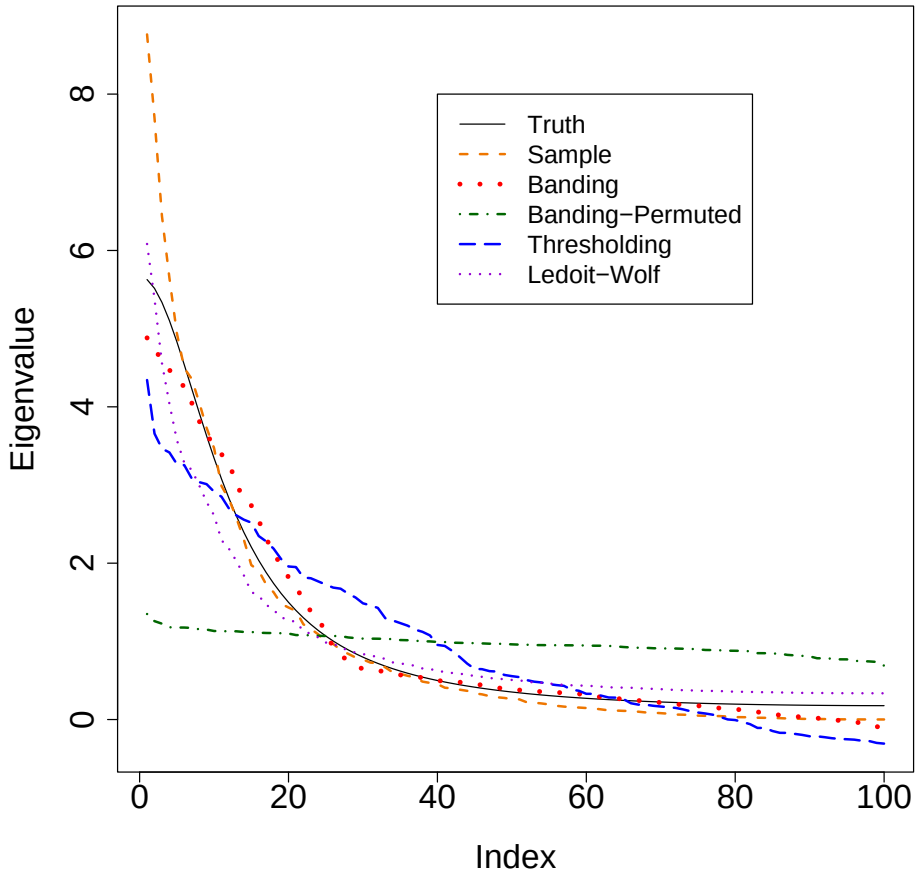


FIG 2. *Scree Plot of Single Realization*

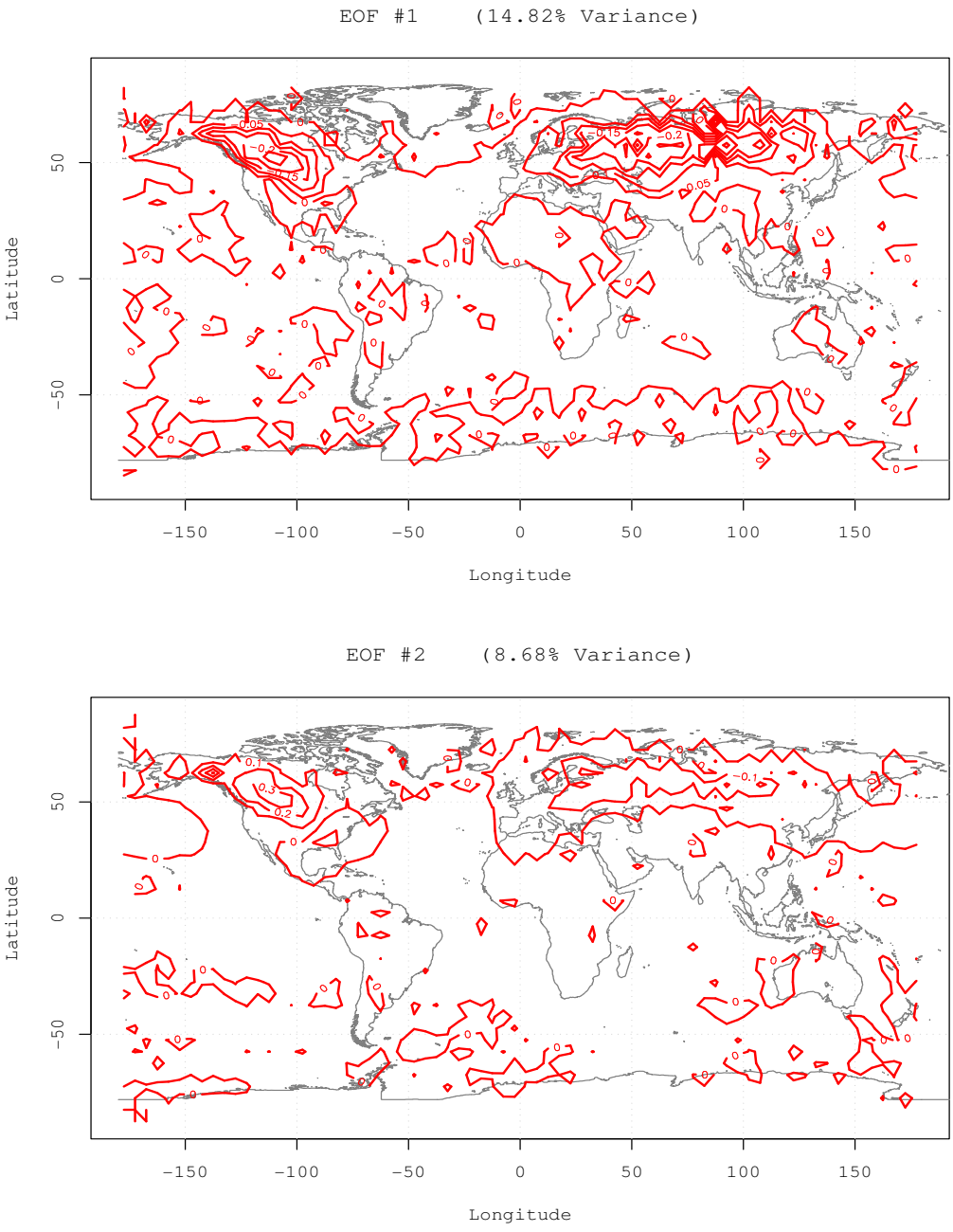


FIG 3. *First two EOFs for the January temperature data obtained from regular PCA.*

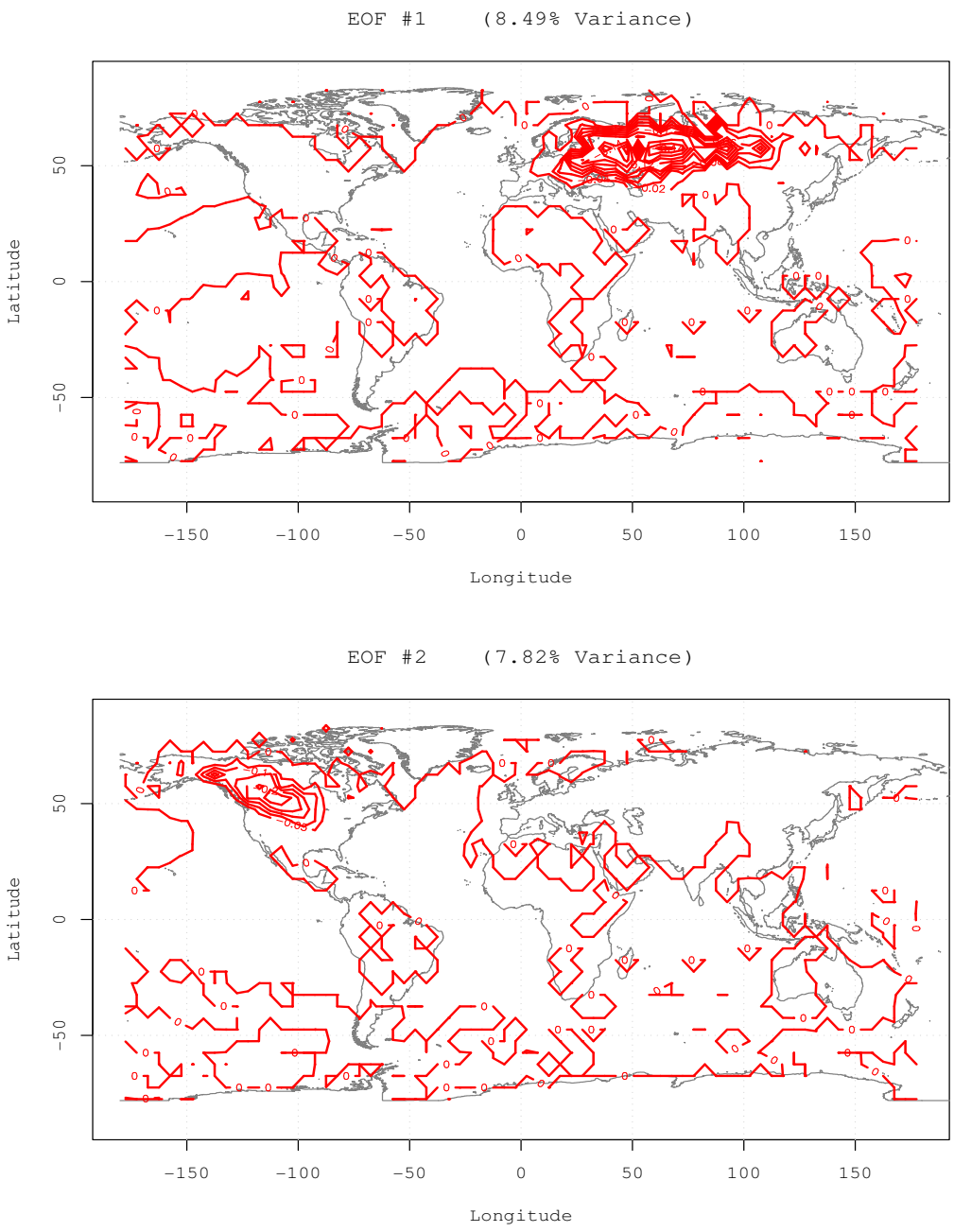


FIG 4. *First two EOFs for the January temperature data obtained from PCA on the thresholded covariance matrix.*