Yu, B. (1997). Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam.* (Edited by D. Pollard, E. Torgersen, and G. Yang eds), pp.423-435. Springer-Verlag.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal components analysis. *J. Comput. Graph. Statist.* **15**, 265-286.

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

E-mail: tcai@wharton.upenn.edu

Department of Statistics, Yale University, New Haven, CT 06511, USA.

E-mail: huibin.zhou@yale.edu

(Received November 2010; accepted July 2011)

COMMENT

Lingzhou Xue and Hui Zou

University of Minnesota

We would like to first congratulate Professors Cai and Zhou for their path-breaking contributions to high-dimensional covariance matrix estimation. Their work greatly deepens our understandings about the nature of large covariance matrix estimation. The technical ideas developed in their work are very useful for studying many high-dimensional learning problems.

1. Geometric Decay Spaces

Throughout this discussion we assume $\log(p) \ll n < p$ and the loss function is the matrix ℓ_1 norm. In their paper, Cai and Zhou (2012) have shown that thresholding is minimax optimal for estimating Σ over the weak ℓ_q ball

$$\mathcal{G}_q(\rho, c_{n,p}) = \{\Sigma : \max_{1\le j\le p} |\sigma_{[k]j}|^q \le c_{n,p}k^{-1},\ \forall\ k,\ \text{and}\ \max_i \sigma_{ii} \le \rho,\ 0 \le q < 1\},$$

and that tapering/banding is minimax optimal for estimating Σ over

$$\mathcal{H}_\alpha(\rho, M) = \{\Sigma : |\sigma_{ij}| \le M|i-j|^{-(\alpha+1)} \text{ for } i \ne j \text{ and } \max_i \sigma_{ii} \le \rho\}.$$

Beyond the polynomial decay space, it is natural to consider covariance matrices with a geometric decay rate. We introduce the parameter spaces

$$\mathcal{A}_\eta(\rho, M) = \{\Sigma : |\sigma_{ij}| \le M\eta^{|i-j|} \text{ for } i \ne j \text{ and } \max_i \sigma_{ii} \le \rho\},$$

$$\mathcal{B}_\eta(\rho, M) = \{\Sigma : \max_{1\le j\le p} |\sigma_{[k]j}| \le M\eta^k, \forall k \text{ and } \max_i \sigma_{ii} \le \rho\},$$

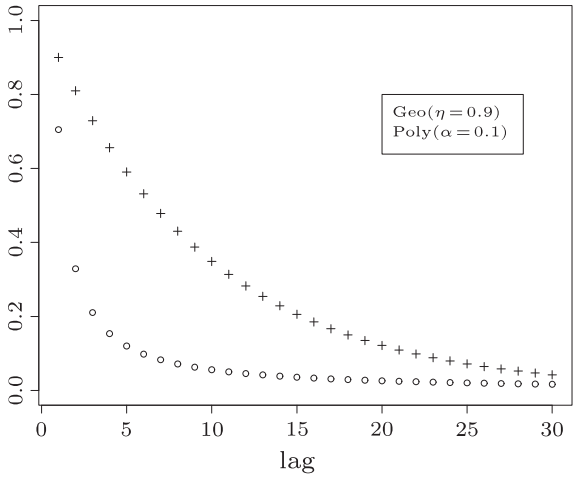

Figure A.1. A slow geometric decay curve versus a slow polynomial decay curve.

where $0 < \eta < 1$. The popular autoregressive matrices belong to geometric decay spaces such as $\mathcal{A}_\eta(\rho, M)$ and $\mathcal{B}_\eta(\rho, M)$. Figure A.1 compares a slow geometric decay curve (with $\eta = 0.9$) and a slow polynomial decay curve (with $\alpha = 0.1$). It is interesting to see that the geometric decay curve is well above the polynomial curve. The reason is that in the polynomial decay case the constant M should be less than 0.705 in order to keep the covariance matrix positive definite. This example suggests that the geometric decay space deserves some special consideration. An important technical contribution in Cai and Zhou (2012) is their carefully designed least favorable distributions for establishing the minimax lower bounds. We follow their idea and give minimax rates under the ℓ_1 norm for geometric decay spaces.

Theorem 1. *Thresholding attains the minimax risk of estimating* Σ *under the matrix* ℓ_1*-norm over* $\mathcal{B}_\eta(\rho, M)$*, with minimax rate*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{B}_\eta(\rho,M)} \mathrm{E}\big\|\hat{\Sigma} - \Sigma\big\|_1^2 \asymp \frac{\log p}{n} \cdot \log^2\Big(\frac{n}{\log p}\Big). \tag{A.1.1}$$

Tapering and banding both attain the minimax risk of estimating Σ *under the* ℓ_1*-norm over* $\mathcal{A}_\eta(\rho, M)$*, with minimax rate*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{A}_\eta(\rho,M)} \mathrm{E}\big\|\hat{\Sigma} - \Sigma\big\|_1^2 \asymp \frac{\log p}{n} \cdot \log\Big(\frac{n}{\log p}\Big). \tag{A.1.2}$$

Theorem 1 also indicates that thresholding is nearly minimax optimal for estimating Σ over $\mathcal{A}_\eta(\rho, M)$. To see that, for $\Sigma \in \mathcal{A}_\eta(\rho, M)$ we have

$$|\sigma_{[k]j}| = \min_{1\le i\le k} |\sigma_{[i]j}| \le \min_{1\le i\le k} M\eta^{|[i]-j|} \le M\eta^{k/2} = M\sqrt{\eta}^k,$$

which immediately implies that $\mathcal{A}_\eta(\rho, M) \subset \mathcal{B}_{\sqrt{\eta}}(\rho, M)$. By Theorem 1 we know that the thresholding estimator achieves a rate of convergence of $(\log p/n) \cdot \log^2(n/\log p)$ over $\mathcal{A}_\eta(\rho, M)$. This rate of convergence differs from the exact minimax lower bound by the factor $\log(n/\log p)$.

The above nearly minimax result is not true in the polynomial decay spaces. Note that $\mathcal{H}_\alpha(\rho, M) \subset \mathcal{G}_{1/(\alpha+1)}(\rho, 2M^{1/(\alpha+1)})$. If we apply the thresholding estimator to estimate Σ over $\mathcal{H}_\alpha(\rho, M)$, the rate of convergence is $(\log p/n)^{\alpha/(1+\alpha)}$. Comparing it to the minimax rate over $\mathcal{H}_\alpha(\rho, M)$ given in Theorem 1 of Cai and Zhou (2012), we see that tapering/banding is fundamentally better than thresholding for estimating bandable matrices over a polynomial decay space.

2. Double Thresholding?

Thresholding estimator is permutation-invariant, whereas banding/tapering estimator requires a natural ordering among variables. It is of interest to combine the strengths of banding and thresholding. This motivates us to consider the double thresholding estimator $\widehat{\Sigma}_{double} = (\hat{\sigma}_{ij}^{double})_{p\times p}$ by performing the entry-wise double thresholding rule

$$\hat{\sigma}_{ij}^{double} = \tilde{\sigma}_{ij}^{block} \cdot I(|\tilde{\sigma}_{ij}^{block}| \geq \lambda_2), \tag{A.2.1}$$

where

$$\tilde{\sigma}_{ij}^{block} = \hat{\sigma}_{ij} \cdot I\big(\max_{(s,t):s-t=i-j} |\hat{\sigma}_{st}| \geq \lambda_1\big).$$

We conducted a small simulation study to compare five regularized covariance matrix estimators (banding, tapering, simple thresholding, block thresholding, and double thresholding). In the simulation study, we considered four covariance models.

- Model 1: $\sigma_{ij} = (1 - |i-j|/\gamma)_+$ for $\gamma = 0.05p$.
- Model 2: $\sigma_{ij} = s_{ij} \cdot (s_{ii}s_{jj})^{-1/2}$, where $S = (I_{p\times p} + U)^T(I_{p\times p} + U) = (s_{ij})_{p\times p}$ with U being a sparse matrix with exactly κ nonzero entries equal to $+1$ or -1 with equal probability for $\kappa = p$.
- Model 3: $\sigma_{ij} = I_{\{i=j\}} + a_{ij}(1+\epsilon)^{-1/2} \cdot I_{\{i\neq j\}}$, where a_{ij} is equal to 0 or $0.6 \cdot |i-j|^{-1.3}$ with equal probability, and ϵ is chosen to be the absolute value of the minimal eigenvalue of $(I_{\{i=j\}} + a_{ij} \cdot I_{\{i\neq j\}})_{p\times p}$ plus 0.01.
- Model 4: $\sigma_{ij} = I_{\{i=j\}} + (1+\epsilon)^{-1/2} \cdot (b_{ij} \cdot I_{\{0<|i-j|\geq 0.5p\}} + c_{ij} \cdot I_{\{|i-j|>0.5p\}})$, where b_{ij} or c_{ij} equals 0 with probability 0.7 and equals $0.7^{|i-j|}$ or $0.7^{|i-j-0.5p|}$ with probability 0.3, and ϵ is chosen to be the absolute value of the minimal eigenvalue of $(I_{\{i=j\}} + b_{ij} \cdot I_{\{0<|i-j|\geq 0.5p\}} + c_{ij} \cdot I_{\{|i-j|>0.5p\}})_{p\times p}$ plus 0.01.

Table A.1. Comparison of banding, tapering, simple thresholding, block thresholding and double thresholding estimators. The standard errors are also shown in the bracket.

	Model 1	Model 2	Model 3	Model 4
Banding	5.55	4.60	1.98	2.15
	(0.08)	(0.00)	(0.01)	(0.01)
Tapering	5.66	4.60	1.98	2.19
	(0.10)	(0.00)	(0.01)	(0.01)
Simple Thresholding	10.61	3.38	2.19	2.40
	(0.19)	(0.03)	(0.02)	(0.02)
Block Thresholding	5.66	4.60	1.91	1.97
	(0.08)	(0.00)	(0.01)	(0.01)
Double Thresholding	5.68	3.38	1.87	1.82
	(0.08)	(0.03)	(0.01)	(0.02)

For each model we generated a training data set with $n = 100$ and $p = 500$ to construct the five estimators, and we also generated an independent validations set of size 100 to tune each estimator. The procedure was repeated 100 times. The estimation accuracy was measured by the matrix ℓ_1 norm averaged over 100 replications. The simulation results are summarized in Table A.1. Model 1 is designed for banding/tapering, and simple thresholding fails miserably there, while blockwise thresholding works as well as tapering. Model 2 is designed for thresholding, and it has a total of 1546 nonzero off-diagonal entries. Banding and tapering fail, as does blockwise thresholding. Model 3 and Model 4 are more interesting examples, because neither banding/tapering nor simple thresholding can give the best estimation. Blockwise thresholding does better than banding/tapering and simple thresholding. However, the best results are given by double thresholding: it significantly outperforms the other four estimators. This simulation study suggests that the double thresholding estimator deserves a more thorough theoretical investigation.

Appendix

For the sake of completeness we give the proof of Theorem 1.

Proof of Theorem 1. We first establish the upper bounds. Recall the thresholding estimator as defined in Cai and Zhou (2012), $\hat{\sigma}_{ij} = \sigma^*_{ij} \cdot I_{\{|\sigma^*_{ij}| \geq \gamma\sqrt{\log p/n}\}}$, where γ is chosen such that $\Pr(|\sigma^*_{ij} - \sigma_{ij}| > \gamma\sqrt{\log p/n}) \leq C_1 p^{-8}$. There exists some integer k^* such that $M\eta^{k^*} > \gamma\sqrt{\log p/n} \geq M\eta^{k^*+1}$. Then we have

$$\sum_i \min\left\{|\sigma_{ij}|, \gamma\sqrt{\frac{\log p}{n}}\right\} \leq k^* \cdot \gamma\sqrt{\frac{\log p}{n}} + M\sum_{i>k^*}\eta^i \leq C\sqrt{\frac{\log p}{n}}\log\left(\frac{n}{\log p}\right).$$

Applying (3.6) and $E\|D\|_1^2 = O(1/n)$ as in Cai and Zhou (2012) yields

$$\sup_{\mathcal{B}_\eta(\rho,M)} E\|\hat{\Sigma} - \Sigma\|_1^2 \le C\left[\frac{\log p}{n}\log^2(\frac{n}{\log p}) + \frac{1}{n}\right] \le C\frac{\log p}{n}\cdot\log^2(\frac{n}{\log p}).$$

For the tapering estimator, we can use the steps for proving Theorem 5 in Cai and Zhou (2012) to obtain

$$\sup_{\mathcal{A}_\eta(\rho,M)} \mathrm{E}\big\|\hat{\Sigma}_k - \Sigma\big\|_1^2 \le C\frac{k^2 + k\log p}{n} + C\frac{\eta^k}{(1-\eta)^2}.$$

Therefore the tapering estimator with $k = \log(n/\log p)/\log(1/\eta)$ can have the rate of convergence

$$\sup_{\mathcal{A}_\eta(\rho,M)} \mathrm{E}\big\|\hat{\Sigma}_k - \Sigma\big\|_1^2 \le C\frac{\log p}{n}\cdot\log(\frac{n}{\log p}).$$

We now prove the lower bounds. Let $\mathcal{H} = \{H_{1,k}, H_{2,k}, \cdots, H_{m_*,k}\}$ be the collection of symmetric matrices with exactly k elements equal to 1 in the first row/column and the rest zeros. To show (A.1.1) we consider

$$\mathcal{B}_0 = \{\Sigma_0 = \rho\cdot I_p \text{ and } \Sigma_m = \rho\cdot I_p + a\cdot H_{m,k} : 1 \le m \le m_*\},$$

where $k = \lfloor \log(n/\log p)/2\log(1/\eta)\rfloor$ and $a = \sqrt{\tau\log p/n}$ for some small constant τ. Since $a \le \eta^k$ still holds, $\mathcal{B}_0$ is a subclass of $\mathcal{B}_\eta(\rho, M)$. Note that $\mathcal{B}_0$ is similar to the space defined in (2.2) in Cai and Zhou (2012) but with a differen k value. Then by Le Cam's lemma and arguments in Section 2.1 of Cai and Zhou (2012), we can show that

$$\sup_{0\le m\le m_*} E\|\hat{\Sigma} - \Sigma_m\|_1^2 \ge \frac{1}{2}\|\mathcal{P}_0 \wedge \bar{\mathcal{P}}\| \cdot \inf_{1\le m\le m_*} \|\Sigma_m - \Sigma_0\|_1^2 \ge c\frac{\log p}{n}\cdot\log^2(\frac{n}{\log p}).$$

Thus the lower bound in (A.1.1) is proved.

To show (A.1.2) we consider

$$\mathcal{A}_0 = \left\{\Sigma_m = \rho\cdot I_p + a\cdot B_{m,k} : 0 \le m \le m_* = \left\lfloor\frac{p}{k}\right\rfloor - 1\right\},$$

where $k = \log(n/\log p)/2\log(1/\eta)$, $a = \sqrt{\log p/16nk}$, and $B_{m,k} = (b_{ij})_{1\le i,j\le p}$ with

$$b_{ij} = I_{\{i=m \text{ and } k+1\le j\le m+k-1\}} + I_{\{j=m \text{ and } k+1\le i\le m+k-1\}}.$$

Since $a^2 \le \log p/n = \eta^{2k}$, $a \le \eta^k$ obviously holds. Then it is easy to show that $\mathcal{A}_0$ is a subclass of $\mathcal{A}_\eta(\rho, M)$. Note that $\mathcal{A}_0$ is similar to the space defined in (2.8) in Cai and Zhou (2012), but with a differen k value. Then by Fano's lemma and the arguments in Section 2.2, we can have

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{A}_0(k,a)} \mathrm{E}\big\|\hat{\Sigma} - \Sigma\big\|_1^2 \ge c\cdot\frac{\log p}{n}\cdot\log(\frac{n}{\log p}).$$

Thus the lower bound in (A.1.2) is proved.

References

Cai, T. and Zhou, H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica* **22**, 1319-1378.

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
E-mail: lzxue@stat.umn.edu
School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
E-mail: zouxx019@umn.edu

(Received March 2012; accepted March 2012)

COMMENT

Tingni Sun and Cun-Hui Zhang

Rutgers University

The estimation of covariance matrices and their inverses is a problem of great practical value and theoretical interest. We congratulate the authors for making an important contribution to it by finding the rate of minimax risk with the ℓ_1 operator norm as the loss function.

A natural question arising from this interesting paper is the minimax rate when the loss is the ℓ_w operator norm $\|M\|_w = \max_{\|u\|_w=1} \|Mu\|_w$. For $\mathcal{G}_q(\rho, c_{n,p})$, the minimax rate has already been established in Cai and Zhou (2012). For $\mathcal{A} = \mathcal{F}_\alpha(\rho, M)$ or $\mathcal{A} = \mathcal{H}_\alpha(\rho, M)$, the upper bound for $w \in [1, 2]$,

$$\inf_{\widehat{\Sigma}} \sup_{\mathcal{A}} \mathrm{E}\big\|\widehat{\Sigma} - \Sigma\big\|_w^2 \lesssim \min\Big\{n^{-2\alpha/(2\alpha+2/w)} + \Big(\frac{\log p}{n}\Big)^{2\alpha/(2\alpha+2/w-1)}, \frac{p^{2/w}}{n}\Big\}, \quad \text{(B.1)}$$

follows from the $\|\cdot\|_2$ bound on the variance term of the tapering estimator and the $\|\cdot\|_1$ bound on the bias term, since $\|M\|_w \le \min\big(\|M\|_1, k^{1/w-1/2}\|M\|_2\big)$ for symmetric $M \in \mathbb{R}^{k\times k}$. Since Σ is symmetric, (B.1) is also valid with w replaced by $w/(w-1) \in [2, \infty]$. For $w = 1$, this gives the minimax rate of the authors. However, it is unclear if the lower bound argument works for $w \in (1, 2)$.

Recent advances in high-dimensional data have been focused on the estimation of high-dimensional objects. However, the estimation of low-dimensional functionals of high-dimensional objects is also of interest. A rate minimax estimator of a high-dimensional parameter does not automatically yield rate minimax estimates of its low-dimensional functionals. For example, instead of the entire