Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.

E-mail: rigollet@princeton.edu

Laboratoire de Statistique, CREST-ENSAE, 3, av. Pierre Larousse, F-92240 Malakoff Cedex, France.

E-mail: Alexandre.Tsybakov@ensae.fr

(Received April 2012; accepted April 2012)

COMMENT

Peter J. Bickel[1], Elizaveta Levina[2], Adam J. Rothman[3] and Ji Zhu[2]

[1] *University of California, Berkeley,* [2] *University of Michigan and* [3] *University of Minnesota*

The authors offer insightful results on minimax rates for large covariance matrix estimation under the matrix ℓ_1-norm that add to the previously known results on the matrix ℓ_2-norm. Incidentally, we expect that some version of the results on the ℓ_1 and ℓ_2 norms in this context can also be developed for the Wiener norm (see Bickel and Lindner (2011) for more details), defined by $\|\Sigma\|_W = \max_k \sum\{|\sigma_{ij}| : |i-j| = k\}$, particularly in the time series domain for which it was introduced by Wiener.

Minimax risk is often used as a benchmark for the evaluation of an estimation method, and having optimal tuning parameter rates is helpful for understanding the behavior of various methods. However, there is also the issue of selecting the tuning parameter in practice, mentioned in the paper as well, which cannot be done using the theoretical bounds of this kind and requires cross-validation. Since this paper studies the convergence in the matrix ℓ_1-norm, and most of the previous literature focuses on convergence in the matrix ℓ_2-norm, we decided to investigate the effect of using various norms for tuning parameter selection via cross-validation, focusing on the thresholding estimator and the parameter space $\mathcal{P}(\mathcal{G}_q(\rho, c_{n,p}))$. Our expectation was that the empirical risk calculated via a particular norm would be minimized by the tuning parameter selected by cross-validation using the same norm, but this turned out not to be the case.

Specifically, we evaluated the performance of the random splitting method for tuning parameter selection described in Bickel and Levina (2008a,b). The n observations are randomly partitioned M times into a validation set of size $n_{\mathrm{va}} = n/\log n$ and a training set of size $n_{\mathrm{tr}} = n - n_{\mathrm{va}}$. Define the ℓ_1-norm

empirical risk $\hat{R}_1$, ℓ_2-norm empirical risk $\hat{R}_2$, and Frobenius norm empirical risk $\hat{R}_F$ as follows:

$$\hat{R}_1(\lambda) = \frac{1}{M}\sum_{m=1}^{M} \|\hat{\Sigma}_\lambda^{(\mathrm{tr,m})} - \hat{\Sigma}^{(\mathrm{va,m})}\|_1,$$

$$\hat{R}_2(\lambda) = \frac{1}{M}\sum_{m=1}^{M} \|\hat{\Sigma}_\lambda^{(\mathrm{tr,m})} - \hat{\Sigma}^{(\mathrm{va,m})}\|_2,$$

$$\hat{R}_F(\lambda) = \frac{1}{M}\sum_{m=1}^{M} \|\hat{\Sigma}_\lambda^{(\mathrm{tr,m})} - \hat{\Sigma}^{(\mathrm{va,m})}\|_F^2,$$

where $\hat{\Sigma}_\lambda^{(\mathrm{tr,m})}$ is the sample covariance computed from the training set of the m-th split and thresholded at λ, and $\hat{\Sigma}^{(\mathrm{va,m})}$ is the sample covariance computed from the validation set of the m-th split.

We generated an i.i.d. sample of size n from $N_p(0,\Sigma)$, where Σ has entries $\sigma_{ij} = 0.4\cdot I(|i-j|=1) + I(i=j)$. Then we selected the tuning parameters $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\lambda}_F$ by minimizing the empirical risks $\hat{R}_1(\lambda)$, $\hat{R}_2(\lambda)$, $\hat{R}_F(\lambda)$, respectively. For each norm, we also computed the "oracle" tuning parameter $\hat{\lambda}_0 = \arg\min_\lambda \|\hat{\Sigma}_\lambda - \Sigma\|$. The performance of each of the tuning parameters was evaluated using the squared L_1 risk, the squared L_2 risk and the squared Frobenius risk, defined respectively as

$$\widehat{\mathrm{E}}\|\hat{\Sigma}_{\hat{\lambda}} - \Sigma\|_1^2, \quad \widehat{\mathrm{E}}\|\hat{\Sigma}_{\hat{\lambda}} - \Sigma\|_2^2, \quad \text{and} \quad \widehat{\mathrm{E}}\|\hat{\Sigma}_{\hat{\lambda}} - \Sigma\|_F^2\, p^{-1},$$

where $\widehat{\mathrm{E}}$ is the average over simulation replications.

We considered two scenarios, $n < p$ and $n > p$. In the $n < p$ scenario, we set $n = p/2$, where p=30, 50, 100, 200 and 500. We used M=10 random splits to estimate the empirical risk and a 200 point resolution for λ. We performed 500 independent replications for $p \le 50$ and 100 independent replications for $p \ge 100$. In the $n > p$ scenario, everything was the same, except for n=60, 100, 200, 500, 1,000 and $p = n/4$.

In Figures D.1 (for $n < p$) and D.2 (for $n > p$) we plot the estimated empirical risks. Each plot corresponds to one evaluation criterion, and the curves on each plot correspond to different methods of selecting the tuning parameter. Surprisingly, the Frobenius norm tuning is always the closest to the oracle, regardless of the evaluation criterion. This is quite counter-intuitive as one would expect, and as was also argued in the paper, that for different evaluation criteria the optimal threshold should be different. Interestingly, however, the Frobenius norm cross-validation tuning is the only one that was analyzed theoretically, in Bickel and Levina (2008b). We may be observing a finite sample phenomenon, but it would be interesting to connect this practical observation to the authors' results on optimal thresholds.

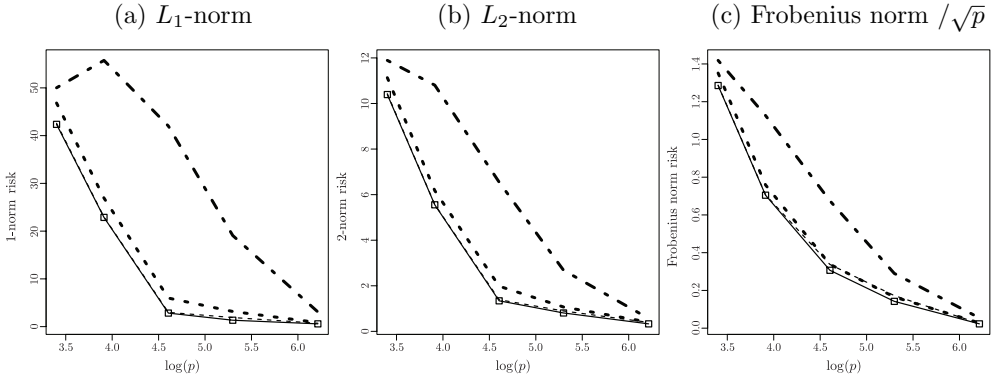

Figure D.1. The $n < p$ scenario. Simulated risk for hard thresholding of the sample covariance matrix with the threshold parameter $\hat{\lambda}_0$ (solid), $\hat{\lambda}_1$ (dots), $\hat{\lambda}_2$ (dash-dot), and $\hat{\lambda}_F$ (dashes).

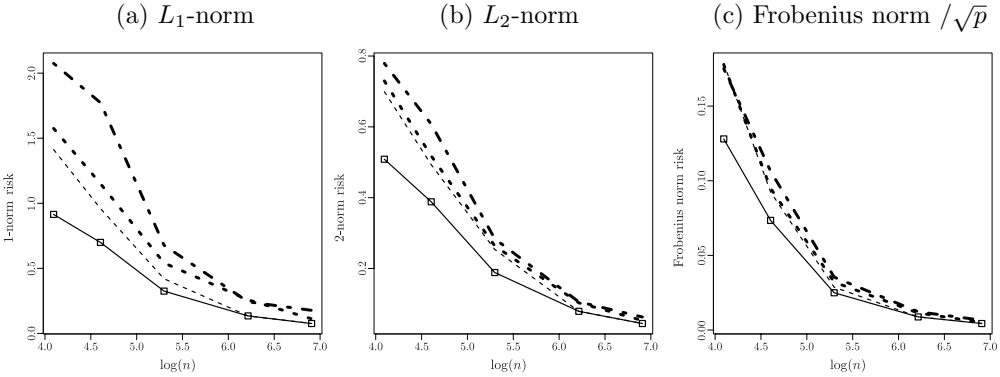

Figure D.2. The $n > p$ scenario. Simulated risk for hard thresholding of the sample covariance matrix with the threshold parameter $\hat{\lambda}_0$ (solid), $\hat{\lambda}_1$ (dots), $\hat{\lambda}_2$ (dash-dot), and $\hat{\lambda}_F$ (dashes).

References

Bickel, P. J. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.

Bickel, P. J. and Levina, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577-2604.

Bickel, P. J. and Lindner, M. (2011). Approximating the inverse of banded matrices by banded matrices with applications to probability and statistics. *Theory Probab. Appl.* **56**, 1-20.

Department of Statistics, University of California, Berkeley, CA 94710-3860, U.S.A.
E-mail: bickel@stat.berkeley.edu

Department of Statistics, University of Michigan, 459 West Hall, Ann Arbor, MI 48109-1107, U.S.A.
E-mail: elevina@umich.edu

School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street, SE Minneapolis, MN 55455, U.S.A.
E-mail: arothman@umn.edu
Department of Statistics, University of Michigan, 459 West Hall, Ann Arbor, MI 48109-1107, U.S.A.
E-mail: jizhu@umich.edu

(Received April 2012; accepted April 2012)

COMMENT

Wei Biao Wu

University of Chicago

I congratulate Professor Cai and Professor Zhou for their timely and important contribution of sharp minimax convergence rates for estimating large covariance matrices. The argument for proving the lower bound is quite sophisticated and is of independent interest. As a useful property, for a class of sparse covariance matrices (cf $\mathcal{G}_q(\rho, c)$ in their (1.1)), the well-known thresholded covariance matrix estimate of Bickel and Levina (2008b) can achieve the minimax rate, while for a class of covariance matrices with weakly correlations (cf $\mathcal{F}_\alpha(\rho, M)$ in (1.2) and $\mathcal{H}_\alpha(\rho, M)$ in (1.3)), a tapered estimate can also have the minimax rate. The paper provides, in the minimax sense, a rigorous justification of the use of the thresholded and the tapered covariance matrix estimates.

My primary concern is the time series application of the large-p-small-n results from the multivariate setting of independent and identically distributed p-variate random vectors. In many time series applications, one has only one realization, $n = 1$. This covariance matrix estimation problem has been discussed by Wu and Pourahmadi (2009), McMurry and Politis (2010), Bickel and Gel (2011), and Xiao and Wu (2012). With $n = 1$, structural assumptions such as stationarity are needed so that the covariance matrix is estimable. Here we propose a possible link between these two settings via block sampling (Politis, Romano, and Wolf (1999)). With observations $X_1, \ldots, X_p$ from a stationary process $(X_i)_{i\in\mathbb{Z}}$, we can consider the $l = \lfloor p/b \rfloor$ blocks $\mathbf{X}_1 = (X_1, \ldots, X_b)'$, $\mathbf{X}_2 = (X_{b+1}, \ldots, X_{2b})'$, $\ldots$, $\mathbf{X}_l = (X_{(l-1)b+1}, \ldots, X_{lb})'$, with b the block size. Consider the estimation of Σ_b, the $b \times b$ covariance matrix of $\mathbf{X}_1$. Assuming weak dependence, one would expect that results similar to (1.5) in their paper can hold.