Discussion of *Large Covariance Estimation by Thresholding Principal Orthogonal Complements* by Fan, Liao and Mincheva

Yi Yu and Richard J. Samworth

February 12, 2013

We congratulate the authors on their paper. POET elegantly tackles low rank plus sparse matrix estimation, provided the eigenvalues of the low rank matrix grow at rate $O(p)$ (see Assumption 1). Suppose now that this assumption does not hold, and instead, we have the following condition.

Assumption 1'. *All the eigenvalues of the $K \times K$ matrix $p^{-\alpha}\boldsymbol{B}'\boldsymbol{B}$ are bounded away from both 0 and ∞ as $p \to \infty$, where $0 < \alpha < 1$.*

Similar conditions are widely used in sparse PCA and low rank plus sparse matrix estimation problems; see, for example, Amini and Wainwright (2009), Agarwal *et al.* (2012). In the following, we consider the three main objectives in Section 2. The notation and model are the same as those in the paper.

Proposition 1' & 2' *Assume Assumption 1. For the factor model with condition (2.1), we have*

$$|\lambda_j - \|\tilde{\boldsymbol{b}}_j\|^2| \leq \|\boldsymbol{\Sigma}_u\|, \quad \textit{for } j \leq K,$$
$$|\lambda_j| \leq \|\boldsymbol{\Sigma}_u\|, \quad \textit{for } j > K.$$

Moreover, if $\{\|\tilde{\boldsymbol{b}}_j\|\}_{j=1}^K$ are distinct, then

$$\|\boldsymbol{\xi}_j - \tilde{\boldsymbol{b}}_{\boldsymbol{j}}/\|\tilde{\boldsymbol{b}}_{\boldsymbol{j}}\|\| = O(p^{-\alpha}\|\boldsymbol{\Sigma}_u\|), \quad \textit{for } j \leq K.$$

From this we see that under a suitable sparsity condition on $\boldsymbol{\Sigma}_u$, the first K principal components are still approximately the same as the columns of the factor loadings, even if the eigenvalues are not as spiked as $O(p)$.

However, for POET to control the relative error of the matrix estimate, Assumption 1 is necessary, as can be seen from a close inspection of the proof of Theorem 2 of Bai and Ng (2002). In fact, if Assumption 1 is replaced with Assumption 1', we have, for $K' < K$, that

$$\lim_{p,T\to\infty} \mathbb{P}\{IC(K') < IC(K)\} > 0.$$

The other half of this theorem still holds, however, so the less spiked structure will not asymptotically increase the risk of over-estimation in the selection of K.

Table 1: For the same $\boldsymbol{u}$ and $\boldsymbol{\mu}_B$ as in Section 6.2, define $\tilde{\boldsymbol{\mu}}_B' = (\boldsymbol{\mu}_B', \boldsymbol{\mu}_B')'$ and expand $\boldsymbol{\Sigma}_B$ to a block diagonal matrix $\tilde{\boldsymbol{\Sigma}}_B$ by making $\boldsymbol{\Sigma}_B$ the diagonal block of $\tilde{\boldsymbol{\Sigma}}_B$. The rows of $\boldsymbol{B}_1$ are generated from a $\mathcal{N}_6(\tilde{\boldsymbol{\mu}}_B, \tilde{\boldsymbol{\Sigma}}_B)$ distribution. Expand the generating process of $\boldsymbol{F}$ similarly to match $\boldsymbol{B}_1$ and generate $\boldsymbol{F}_1$ accordingly, and then let $\boldsymbol{Y} = C_1\boldsymbol{B}_1\boldsymbol{F}_1' + \boldsymbol{u}$. Here, $K = 6$, $K_{\max} = 20$. The means of the estimated K are reported over 100 repetitions, with standard errors in brackets.

Methods	$C_1 = 1$	$C_1 = 1/3$	$C_1 = 1/10$	$C_1 = 10$
IC	6.00(0.00)	1.08(0.27)	1.00(0.00)	6.00(0.00)
AIC	20.00(0.00)	20.00(0.00)	20.00(0.00)	20.00(0.00)
BIC	6.00(0.00)	2.00(0.00)	1.00(0.00)	6.00(0.00)

The performances of IC, AIC and BIC are compared in Table 1, with the corresponding largest eigenvalues of $\boldsymbol{Y}\boldsymbol{Y}'$ in Figure 1. If the spectrum structure satisfies Assumption 1 ($C_1 \geq 1$), both IC and BIC select the correct value of K. However, if we shrink the spiked eigenvalues, IC and BIC tend to underestimate, while AIC overestimates, the true K.

To examine the effect of missing the Kth common factor, assume (2.1) and that $\text{rank}(\boldsymbol{B}'\boldsymbol{B}) = K$, but the estimator is

$$\hat{\boldsymbol{\Sigma}}_{K-1} = \Sigma_{i=1}^{K-1}\hat{\lambda}_i\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i' + \hat{\boldsymbol{R}}_{K-1}^{\mathcal{T}},$$

where $\hat{\boldsymbol{R}}_{K-1}^{\mathcal{T}}$ is the entrywise-shrunk estimator of $\boldsymbol{R}_{K-1} = \boldsymbol{b}_K\boldsymbol{f}_K\boldsymbol{f}_K'\boldsymbol{b}_K' + \boldsymbol{\Sigma}_u$. In this case, due to the common factor, most of the pairs of cross-sectional units in $\boldsymbol{R}_{K-1}$ are no longer "weakly correlated". Note that the $\hat{\theta}_{ij}$'s in Appendix A are still the same, i.e., no extra shrinkage is introduced. However, m_p used in Theorem 2 and 3 is not $o(p)$, so the error bound does not converge to zero. On the other hand, when K is correctly or over-estimated, even substituting Assumption 1' for Assumption 1, the corresponding results in Theorems 2 and 3 still hold. Thus, if there is doubt about the validity of Assumption 1, a less severe penalty (e.g. AIC) may be preferable, to avoid the more serious error of underestimation of K.

References

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012) Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *The Annals of Statistics*, **40**, 1171–1197.

Amini, A. A. and Wainwright, M. J. (2009) High-dimensional analysis of semidefinite relaxations for sparse principal components. *The Annals of Statistics*, **37**, 2877–2921.

Bai, J. and Ng, S. (2002) Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191–221.

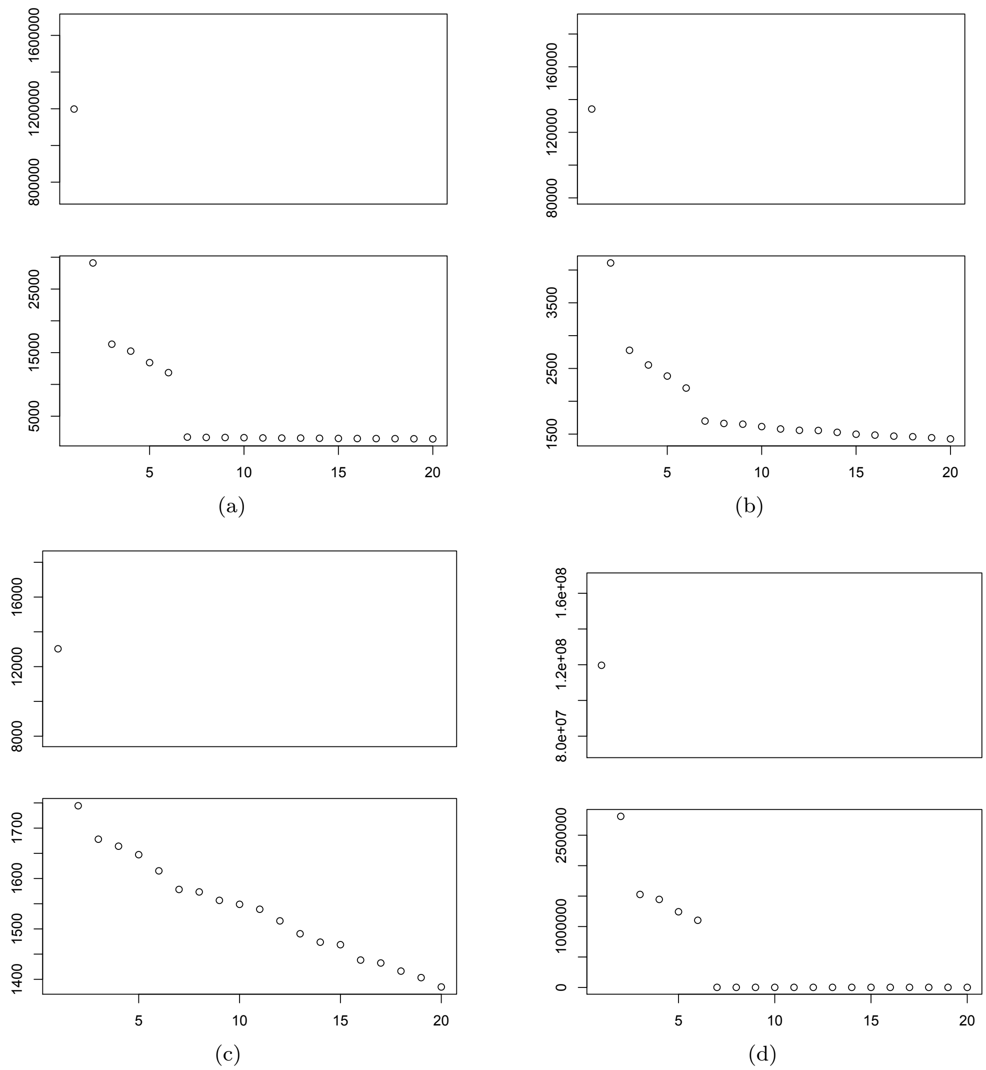

Figure 1: The largest 20 eigenvalues of $\boldsymbol{YY'}$ in cases (a) $C_1 = 1$ (b) $C_1 = 1/3$, (c) $C_1 = 1/10$ and (d) $C_1 = 10$.