Table 12. Comparison of the root finding and the alternating projections procedures†

Procedure	*Computing time (s)*	*Minimum eigenvalue*	*Number of iterations to converge*
POET	2.34 (0.148)	< 0	—
Root finding	62.5 (9.93)	0.149 (0.054)	20.7 (4.14)
Alternating projections	2.43 (0.118)	0.0997 (0.000)	7.91 (0.71)

†Numbers in cells are averages over 100 data sets along with their empirical standard deviations in parentheses. The code is written in Octave 3.2.3 (Eaton, 2002) on a laptop computer (MacBook Air, 1.8 GHz i5 processor with 4 Gbytes memory). Covariance hard thresholding was used in the ordinary POET with $C = 0.1$ and $K = 3$. In the root finding, Octave function `fzero()` was used to find $C_{\min}$, the root of equation (4.1), starting from $C = 0.1$; final thresholding was conducted for $C_{\min} + 0.1$. In the alternating projections, the lower bound μ for the minimum eigenvalue was set to 0.1. Both procedures terminated if $C_{\min}$ or $\lambda_{\min}$ does not change up to the third digit after the decimal point.

We numerically compared two procedures in a simple setting. The comparison was done for 100 data sets with $n = 50$ samples of a $p = 100$-dimensional standard normal vector. The results are summarized in Table 12. The root finding took roughly 1 min to converge, whereas the alternating projections converged in 2.5 s, with little additional time to the ordinary POET-estimator, i.e. POET without adjustment for positive definiteness, in less than half the iterations.

The POET-method may theoretically be optimization free, but the *post hoc* adjustment to make the ordinary POET-estimator positive definite involves some numerical optimization anyway. A little more attention to this step may greatly improve the practicality of the method proposed.

Lingzhou Xue (*Princeton University*) **and Hui Zou** (*University of Minnesota, Minneapolis*)

We first congratulate Fan, Liao and Mincheva for their innovative and timely contribution to high dimensional covariance matrix estimation. POET is a statistically and computationally appealing method for estimating a large covariance matrix with a conditional sparsity structure. We discuss two alternative methods for estimating the error covariance matrix in POET.

POET2 via positive definite adaptive thresholding estimation

POET uses adaptive thresholding estimation (Cai and Liu, 2011) on the principal orthogonal complement $\hat{\mathbf{\Sigma}}_{u,\hat{K}} = (\hat{\sigma}_{ij}^{u,\hat{K}})_{p\times p}$ to estimate the sparse error covariance matrix, namely

$$\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}} = (\hat{\sigma}_{ij}^{u,\hat{K}} I_{\{i=j\}} + s_{ij}(\hat{\sigma}_{ij}^{u,\hat{K}}) I_{\{i\neq j\}})_{p\times p}$$

where $\tau_{ij} = C w_{\mathcal{T}} \sqrt{\hat{\theta}_{ij}} > 0$ is the entry-dependent threshold. In Section 4.1 Fan, Liao and Mincheva discussed the importance of choosing a proper threshold to guarantee the finite sample positive definiteness of $\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}}$. POET chooses the threshold C in the range $(C_{\min} + \varepsilon, M)$ where $C_{\min}$ is defined in expression (4.1). Xue *et al.* (2012) proposed a direct convex programme to deliver a positive definite thresholding covariance matrix estimator. We adopt the idea thereof to construct another positive definite adaptive thresholding estimator for POET. Specifically, we consider the following constrained l_1-minimization problem:

$$\hat{\mathbf{\Sigma}}_{u,\hat{K}}^{\mathcal{T}_2} = \arg\min_{\mathbf{\Sigma}\succeq\varepsilon\mathbf{I}} \frac{1}{2}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_{u,\hat{K}}\|_{\mathrm{F}}^2 + \sum_{(i,j):i\neq j} \tau_{ij}|\sigma_{ij}|,$$

where $\varepsilon > 0$ is some arbitrarily small constant. The alternating direct method of multipliers algorithm in Xue *et al.* (2012) can be easily modified to solve $\hat{\Sigma}_{u,\hat{K}}^{\mathcal{T}_2}$. We introduce a new variable $\mathbf{\Theta}$ and an equality constraint $\mathbf{\Sigma} = \mathbf{\Theta}$, namely

$$(\hat{\mathbf{\Theta}}^{+}, \hat{\mathbf{\Sigma}}^{+}) = \arg\min_{\mathbf{\Theta},\mathbf{\Sigma}} \left\{ \frac{1}{2}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_{\mathrm{F}}^2 + \sum_{(i,j):i\neq j} \tau_{ij}|\sigma_{ij}| : \mathbf{\Sigma} = \mathbf{\Theta}, \mathbf{\Theta} \succeq \varepsilon\mathbf{I} \right\}.$$

We minimize its augmented Lagrangian function for some given parameter $\rho > 0$ (for simplicity we can fix $\rho = 1$), i.e.

$$L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{\Lambda}) = \frac{1}{2}\|\boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_{u,\hat{K}}\|_{\mathrm{F}}^2 + \sum_{(i,j):i\neq j} \tau_{ij}|\sigma_{ij}| - \langle \boldsymbol{\Lambda}, \boldsymbol{\Theta} - \boldsymbol{\Sigma}\rangle + \frac{1}{2\rho}\|\boldsymbol{\Theta} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2$$

We iteratively solve $L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}; \boldsymbol{\Lambda})$ for $(\boldsymbol{\Theta}^{i+1}, \boldsymbol{\Sigma}^{i+1})$ by alternating minimization, and then we update the Lagrange multiplier $\boldsymbol{\Lambda}^{i+1}$. The complete alternating direct method of multipliers algorithm proceeds as follows.

For $i = 1, 2, \ldots,$: $\boldsymbol{\Theta}$ step,

$$\boldsymbol{\Theta}^{i+1} = \arg\min_{\boldsymbol{\Theta}\succeq\varepsilon\mathbf{I}} L(\boldsymbol{\Theta}, \boldsymbol{\Sigma}^i; \boldsymbol{\Lambda}^i) = (\boldsymbol{\Sigma}^i + \rho\boldsymbol{\Lambda}^i)_+;$$

$\boldsymbol{\Sigma}$ step

$$\boldsymbol{\Sigma}^{i+1} = \arg\min_{\boldsymbol{\Sigma}} L(\boldsymbol{\Theta}^{i+1}, \boldsymbol{\Sigma}; \boldsymbol{\Lambda}^i) = \frac{1}{1+\rho}(\mathrm{ST}\{\rho(\hat{\sigma}_{jk}^n - \boldsymbol{\Lambda}_{jk}^i) + \boldsymbol{\Theta}_{jk}^{i+1}, \tau_{jk}\rho\})_{p\times p};$$

$\boldsymbol{\Lambda}$ step

$$\boldsymbol{\Lambda}^{i+1} = \boldsymbol{\Lambda}^i - \frac{1}{\rho}(\boldsymbol{\Theta}^{i+1} - \boldsymbol{\Sigma}^{i+1}).$$

The two operators $(\cdot)_+$ and $\mathrm{ST}(\cdot)$ are defined in Xue *et al.* (2012).

We call $\hat{\boldsymbol{\Sigma}}_{\hat{K}}^{\mathcal{T}_2} = \boldsymbol{\Sigma}_{i=1}^{\hat{K}} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{T}_2}$ the POET2 estimator of $\boldsymbol{\Sigma}$. We compared POET2 and POET by using simulation models 1–3 with $T = 200$ and $p = 200$ from Section 6.5.2. As can be seen from Table 13, the two versions of POET have very similar performance.

POET3 via principal orthogonal complement banding

If $\boldsymbol{\Sigma}_u$ is in fact bandable, another version of POET can use banding instead of thresholding to regularize the principal orthogonal complement. The bandable structure is widely used to model dependence between ordered variables. Given a banding parameter k, principal orthogonal complement banding yields $\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{B}} = (\hat{\sigma}_{ij}^{u,\hat{K}} I(|i-j| \leqslant k))_{p\times p}$. To guarantee the positive definiteness, we consider the eigendecomposition of $\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{B}} \Sigma_i \hat{\lambda}_i \mathbf{v}_i \mathbf{v}_i'$, and then define $\hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{B}+} = \Sigma_i \max(\hat{\lambda}, 0)\mathbf{v}_i \mathbf{v}_i'$. The POET3-estimator of $\boldsymbol{\Sigma}$ is defined as $\hat{\boldsymbol{\Sigma}}_{\hat{K}}^{\mathcal{T}_3} = \Sigma_{i=1}^{\hat{K}} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\boldsymbol{\Sigma}}_{u,\hat{K}}^{\mathcal{B}+}$. We compared POET3 and POET by using simulation models 1 and 2. As shown in Table 14, POET3 performs better than POET by taking advantage of the bandable structure. However, POET3

Table 13. Comparison of POET2 and POET in terms of average spectral norm loss over 100 replications ($T = 200$, $p = 200$)

	Results for model 1		*Results for model 2*		*Results for model 3*	
	POET	*POET2*	*POET*	*POET2*	*POET*	*POET2*
$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$	26.20	26.18	2.04	2.04	7.73	7.74
$\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|$	1.31	1.30	2.07	2.06	8.48	8.50

Table 14. Comparison of POET3 and POET in terms of average spectral norm loss over 100 replications ($T = 200$, $p = 200$)

	Results for model 1		*Results for model 2*	
	POET	*POET3*	*POET*	*POET3*
$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$	26.20	25.76	2.04	1.68
$\|\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\|$	1.31	1.26	2.07	1.73

is potentially better only when the bandable structure is reliable and the ordering information is accurate. Otherwise, POET (or POET2) should be preferred.

The **authors** replied later, in writing, as follows.

We are very grateful to all contributors for their stimulating comments and questions on high dimensional covariance matrix estimation in the presence of common factors. They have touched many important issues, from theoretical understanding to methodological improvements and applications. Their contribution is important for the better understanding of the proposed POET-estimator. We shall not be able to resolve all points in a brief rejoinder. Indeed, the discussion can be seen as a collective research agenda for the future, and some of the agendas have already been undertaken by the discussants.

Spiked eigenvalues

Several discussants (Critchley, Jung and Fine, Lam and Hu, Linton and Vogt, Onatski, and Yu and Samworth) gave their detailed comments and questions regarding the spikiness of the eigenvalues. They express some concern that the separation between large and remaining eigenvalues is too distinct. Their concerns are very relevant. If there are no large gaps between the large eigenvalues and the small ones, the systematic component of the covariance cannot even be differentiated from the idiosyncratic part in our factor model: $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}' + \mathbf{\Sigma}_u$. We impose the pervasiveness of the factors through the assumption that the eigenvalues of the $K \times K$ matrix

$$\mathbf{A}_p \equiv \frac{1}{p}\mathbf{B}'\mathbf{B} = \frac{1}{p}\sum_{i=1}^{p} \mathbf{b}_i \mathbf{b}_i'$$

are bounded away from both 0 and ∞ as p grows. The interpretation of this is very natural: the factors are common to the majority of variables. Under this condition and the sparsity assumption on $\mathbf{\Sigma}_u$, the first K eigenvalues are of order p whereas the remaining eigenvalues are bounded.

This pervasiveness is not the minimum condition to make the problem identifiable. As correctly pointed out by Jung and Fine, the spikiness of the eigenvalues of the low rank matrix $\mathbf{B}\mathbf{B}'$ and sparseness of $\mathbf{\Sigma}_u$ together play an important role in distinguishing the systematic and idiosyncratic components. As long as $\|\mathbf{\Sigma}_u\|$ is much smaller than $\|\mathbf{B}\mathbf{B}'\|$, these two components can be distinguished. Of course, the rates of convergence depend on the size of the gaps and other parameters. For example, Yu and Samworth suggested a weaker version of the pervasive condition, which replaces p^{-1} in the definition of $\mathbf{A}_p$ with $p^{-\alpha}$ for some $\alpha \in (0, 1)$. With this weaker condition, all results should still go through, and carefully inspecting our technical proofs should yield the rates of convergence. In contrast, there is also recent literature that requires $\alpha = 0$ or replaces $p^{-\alpha}$ with $\log(p)^{-1}$, which corresponds to approximately 'sparse loading matrices' (Pati *et al.*, 2012; Carvalho *et al.*, 2009). See also the discussion by Pan and Peng for a novel approach. Intuitively, this allows for non-pervasive (weak) factors that have no effect on a non-negligible portion of the individuals. However, this will bring more difficulty to estimating the number of spiked eigenvalues, and identifying the low rank part from the idiosyncratic part, because the signal is too weak.

We agree wholeheartedly with H. Huang, Y. Liu, Marron, D. Shen and H. Shen that now is a good time to study asymptotic contexts, where the first K eigenvalues of $\mathbf{\Sigma}$ grow quickly. Indeed, sparsity appears rarely in applications, yet conditional sparsity is likely to be more relevant for many applications. Studying spiked eigenvalues amounts to exploring the main structure of the covariance matrix.

We agree on the existence of weaker factors in applications (Lam and Hu, Linton and Vogt, and Onatski). These factors are usually difficult to differentiate from the idiosyncratic components and do not play a noticeable role without a large amount of data. We would like to add that our assumption on the spikiness of eigenvalues is imposed on the population covariance, not on the sample covariance matrix. Model diagnostics based on sample eigenvalues should be interpreted with care owing to large estimation errors in high dimensional matrices.

Choice of the number of factors K

The gaps between the spiked eigenvalues and the remaining eigenvalues have impact on the choice of the number of factors K, Fryzlewicz and N. Huang, Lam and Hu, and other discussants carried out many interesting simulations about the issue of choosing K, the number of these spiked eigenvalues. In many simulations by the contributors, the responses are not driven by a few common factors. In contrast, POET builds on the principal components analysis based on the sample covariance matrix, whose first K eigenvalues are growing at rate $O(p)$. The existence of these spiked eigenvalues is implied by the pervasive condition for the common factors. This gap can be made smaller if Yu and Samworth's assumption is