Marco A. R. Ferreira (*University of Missouri, Columbia*)

I congratulate Professor Fan and his colleagues for their valuable contribution to the area of large covariance matrix estimation.

Professor Fan and colleagues have developed a method for estimating large covariance matrices when there are common unobservable factors and additional cross-sectional correlation. They consider the case when, as the number of individuals p and the number of time points T grow to infinite, the number of common unobservable factors K remains fixed. In addition, in their set-up the eigenvalues corresponding to the common factors are divergent as $p \to \infty$. Finally, they assume that the covariance matrix of the idiosyncratic component is approximately sparse.

With these assumptions, the authors develop a method based on principal component analysis for covariance matrix estimation. Specifically, first they estimate the contribution of the common factors to the covariance matrix by the sum of the K first terms of the sample covariance matrix spectral decomposition. Then, they subtract the estimated common factors contribution from the sample covariance matrix to obtain the principal orthogonal complement. Further, they apply thresholding to the principal orthogonal complement to obtain an estimator of the idiosyncratic covariance matrix. Finally, their covariance matrix estimator is the sum of the estimated common factors contribution and the estimated idiosyncratic covariance matrix.

I have two main comments or questions on the paper.

(a) As the number of individuals p increases, it seems intuitive to assume that the underlying process generating the data should grow in complexity, i.e. it seems intuitive that K should grow with p. What would be the potential technical issues that would arise if one decides to extend the current work to the case when K grows with p?
(b) For the application of thresholding, there are a number of constants that must be chosen such as τ in equation (2.6) and C in equation (3.2). There seems to be an opportunity for the use of empirical Bayes methodology for the estimation of those threshold parameters.

Florian Frommlet (*Medical University Vienna*)

I congratulate the authors on this impressive paper concerned with estimating high dimensional covariance matrices under conditional sparsity. Their approach is surprisingly simple: first compute the principal components of the sample covariance matrix, then estimate the number of relevant components and finally apply a thresholding procedure on the remaining covariance matrix. In spite of this simplicity extensive simulation studies in their paper show that POET, the implementation of the approach presented, outperforms competing algorithms in various scenarios.

It is not too surprising that POET performs well in those scenarios based on factor models with few factors, which mimic the situation under which the authors have derived asymptotic results for their method, i.e. when the covariance matrix has a small (fixed) number K of very large eigenvalues. It is quite intuitive that in this situation the first K principal components will simply represent the corresponding factors of the factor model. Also it appears to be clear that the procedure works well when no factors are present, as long as the number of components is then correctly estimated to be 0.

For me the most astonishing result is that POET appears to do relatively well in model 3 of Section 6.5.2, where data were simulated according to an auto-regressive AR(1) model. This is the only presented simulation scenario where data were not simulated either from a factor model with a small number of strong factors, or alternatively from a model without factors and sparse covariance matrix. The covariance matrix of the suggested AR(1) model does not have particularly spiked eigenvalues, but the eigenvalues smoothly decrease from their maximum. In fact for $p = 200$ and $p = 300$ there are 36 and 53 eigenvalues larger than 1 respectively. According to the simulations presented POET picks for this scenario (both for $p = 200$ and $p = 300$) on average roughly six factors to model the covariance structure, outperforming direct thresholding of the sample covariance matrix. This result indicates that POET might work well even in situations which are not covered by the asymptotic analysis presented. However, further work seems to be necessary to explain why that would be so.

I. Gijbels and K. Herrmann (*KU Leuven*) **and A. Verhasselt** (*Universiteit Hasselt and Universiteit Antwerpen*)

Fan and his colleagues present a very nice estimation technique (POET) for high dimensional covariance estimation, based on principal component estimation and thresholding the orthogonal complement of the principal components. They show that POET is equivalent to constrained least squares (CLS) estimation.

We wonder how robust POET is when the data matrix is corrupted, since it is well known that least-

squares-based methods are not robust. The equivalence of POET to a CLS estimation problem seems to open the way for a more robust procedure. The use of robust principal component methods (e.g. Engelen *et al.* (2005)) could also offer a possibility.

As pointed out in the literature (see for example Antoniadis (2007)), the qualitative properties of a thresholding rule turn out to be important. For example, the hard thresholding rule is discontinuous, whereas the soft thresholding rule is continuous. In CLS regression hard thresholding leads to a larger variance of the estimates, whereas soft thresholding shifts the estimates, creating a bias. What is the effect of such qualitative properties of the thresholding rule on the POET estimator?

The authors use a computationally expensive cross-validation criterion to choose C. It might be worth the effort to exploit the equivalent CLS problem and to use criteria based on this equivalence, such as an Akaike type of criterion.

Portfolio allocation in the Markowitz (1952) framework is chosen as a numerical illustration of POET. In the simulation studies and empirical application, the emphasis is on estimating the weights of the minimum variance (MV) portfolio as the solution to $\mathbf{w}_{\mathrm{MV}}(\boldsymbol{\Sigma}) = \arg\min_{\mathbf{w}'\mathbf{1}=1} \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w}$. This is in line with current literature (Kourtis *et al.* (2012) and references therein) where the MV portfolio is preferred because it alleviates the necessity of estimating the stock returns. As the MV weights admit the expression $\mathbf{w}_{\mathrm{MV}}(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}^{-1}\mathbf{1}/(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1})$, the comparison of POET, the strict factor model (Fan *et al.*, 2008) and the sample covariance (SC) matrix estimator amounts to a comparison of the estimated precision matrix in the models considered. It is known that the SC precision matrix performs poorly (Fan *et al.*, 2008; Kourtis *et al.*, 2012) and measures to counterbalance estimation errors must be taken. In a $p < T$ framework shrinkage methods for example are applied to the SC matrix before (Ledoit and Wolf, 2003) or after inversion (Kourtis *et al.*, 2012), significantly enhancing results. Shrinkage methods have also been applied to the $p \gg T$ framework (see Ledoit and Wolf (2004)), establishing a possible competitor in this scenario as well. Owing to the known shortcomings of the SC precision matrix deeper insights can be expected from a comparison with such refined methods.

Wally Gilks (*University of Leeds*)

Fan and his colleagues state that the low rank plus sparse representation of their model is for the *population* covariance matrix. The most obvious interpretation of this assertion is that the model is intended to describe the population, not the specific individuals sampled. This interpretation is somewhat at odds with the design of the sparse component of the model $\boldsymbol{\Sigma}_u$, which accounts for idiosyncratic correlations between *specific* individuals.

At a population level, such idiosyncratic components can only be represented in terms of *probabilities* of idiosyncratic correlation. For example, suppose that two individuals i and j, randomly and independently sampled from the population, have a probability π of interacting idiosyncratically. Suppose further that the covariance $\sigma_{u,ij}$ of their idiosyncratic errors is ρ if i and j interact, and 0 otherwise. In a sample of size p, the probability that individual i idiosyncratically interacts with k other individuals is distributed as binomial$(\pi, p-1)$. The authors require that their measure of sparsity, $m_p = \max_{i \leqslant p} \Sigma_{j \leqslant p} |\sigma_{u,ij}|^q$, grows with sample size as $o(p)$. Letting $\sigma_{u,ij} = \tau$, we have

$$\begin{aligned} m_p &= \tau^q + \rho^q K_p^{(1)} \\ &> \tau^q + \rho^q \bar{K}_p \\ &\approx \tau^q + \rho^q (p-1)\pi \\ &\geqslant \rho^q \pi p \end{aligned}$$

where $K_p^{(1)}$ and $\bar{K}_p$ denote the maximum and mean values in a sample of size p from a binomial$(\pi, p-1)$ distribution. Thus, $m_p \neq o(p)$ unless $\rho = 0$.

Hajo Holzmann and Anna Leister (*Philipps-Universität Marburg*)

We congratulate Fan and his colleagues for an inspiring paper on estimating the factor structure in high dimensional, approximate factor models, and its consequences for estimating the underlying covariance matrix.

Let us consider implications for the time series structure of $(\mathbf{y}_t)$, specifically its lagged covariance matrix, and convergence in the $\|\cdot\|_{\max}$-norm.

When estimating $\Sigma = \mathrm{cov}(\mathbf{y}_t)$, Fan *et al.* (2011), theorem 3.2, obtain the rate $O_p[\sqrt{\{\log(p)/T\}}]$ in $\|\cdot\|_{\max}$ for an estimate based on an observed factor structure, whereas in the present paper, utilizing estimated factors, the authors obtain the rate $O_p[1/\sqrt{p} + \sqrt{\{\log(p)/T\}}]$. Now, for the sample covariance, writing