Large covariance estimation by thresholding principal orthogonal complements

J. Fan, Y. Lian & M. Mincheva

Journal of the Royal Statistical Society B, 2013

Discussion led by Esther Salazar
Duke University

April 4, 2014

Summary

- This paper deals with the estimation of a high dimensional covariance matrix
- Let y_{it} be the observed response for the ith individual $(i = 1, \ldots, p)$ at time $t = 1, \ldots, T$ such that p is 'large'
- The objective is the estimation of the covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$

- Paper's contribution: The authors develop a method for estimating large covariance matrices by assuming a sparse error covariance matrix in an approximate factor model
- Their method is called *Principal orthogonal complement thresholding* (POET)
- Basically, the POET estimator is the sum of two parts
  (1) the non-sparse low rank part resulting from the factor model
  (2) the sparse part arising as a result of thresholding the "principal orthogonal complement"

On the estimation of large covariance matrices

- In recent years researchers have proposed various regularization techniques to estimate $\mathbf{\Sigma}$
- Key assumption: $\mathbf{\Sigma}$ is sparse (many entries are 0 or nearly so)
- In some applications the sparsity assumption on $\mathbf{\Sigma}$ is not appropriate. Examples: financial returns depend on the equity market risks, housing prices depend on the economic health, ...
- A natural extension is conditional sparsity. Given the common factors, the outcomes are weakly correlated.

High dimensional factor analysis

Consider an approximate factor model

$$\boldsymbol{y}_t = \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{u}_t$$

$y_t = (y_{1t}, \ldots, y_{pt})'$ is the observed response for p individuals at time t, $\boldsymbol{B} \in \mathbb{R}^{p \times K}$ is factor loading matrix, $\boldsymbol{f}_t \in \mathbb{R}^K$ is a vector of common factors and u_t the error term. Both p and T diverge to ∞

• Our interest: $\boldsymbol{\Sigma}$, the $p \times p$ covariance matrix of $\boldsymbol{y}_t$ which is assumed to be time invariant. Also

$$\boldsymbol{\Sigma} = \boldsymbol{B}\text{cov}(\boldsymbol{f}_t)\boldsymbol{B}' + \boldsymbol{\Sigma}_u$$

where $\boldsymbol{\Sigma}_u = (\sigma_{u,ij})_{p \times p}$ is the covariance matrix of $\boldsymbol{u}_t$.
• Literature on approximate factor models typically assumes that the first K eigenvalues of $\boldsymbol{B}\text{cov}(\boldsymbol{f}_t)\boldsymbol{B}'$ diverge at rate $\mathcal{O}(p)$ whereas all the eigenvalues of $\boldsymbol{\Sigma}_u$ are bounded as $p \to \infty$

$$\boldsymbol{\Sigma} = \boldsymbol{B}\text{cov}(\boldsymbol{f}_t)\boldsymbol{B}' + \boldsymbol{\Sigma}_u$$

In this paper the authors assume that $\boldsymbol{\Sigma}_u$ is *approximately sparse* as in Bickel and Levina (2008)[1] and Rothman et al. (2009)[2]: for some $q \in [0, 1)$,

$$m_p = \text{max}_{i \leq p} \sum_{j \leq p} |\sigma_{u,ij}|^q$$

does not grow too fast as $p \to \infty$. If $q = 0$ then $m_p = \text{max}_{i \leq p} \sum_{j \leq p} \boldsymbol{I}_{(\sigma_{u,ij} \neq 0)}$ is the maximum number of non-zero elements

The conditional sparsity structure was already explored by Fan et al(2011)[3] when the factors $\{\boldsymbol{f}_t\}$ are observable, using regression analysis to estimate $\{\boldsymbol{u}_t\}$

Here the factor scores $\{\boldsymbol{f}_t\}$ are unknown and must be inferred

[1] Covariance regularization by thresholding. *Annals of Statistics*

[2] Generalized thresholding of large covariance matrices. *JASA*

[3] High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*

Objectives of this paper

(a) to understand the relationships between PCA and high dimensional factor analysis

(b) to estimate both covariance matrices $\mathbf{\Sigma}$ and the idiosyncratic $\mathbf{\Sigma}_u$

(c) to investigate the effect of estimating the unknown factors on the covariance estimation

The POET method

The proposed approach is simple and optimization free and it uses the data only through the sample covariance matrix $\hat{\mathbf{\Sigma}}_{sam} \in \mathbb{R}^{p \times p}$

They propose a non-parametric estimator of $\mathbf{\Sigma}$ based on PCA. Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$ be the ordered eigenvalues of $\hat{\mathbf{\Sigma}}_{sam}$ and $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^p$ be their corresponding eigenvectors

Algorithm

- Run the singular value decomposition (SVD) on $\hat{\mathbf{\Sigma}}_{sam}$ of $\boldsymbol{y}_t$. Then the sample covariance has the following spectral decomposition

$$\hat{\mathbf{\Sigma}}_{sam} = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\boldsymbol{R}}_K$$

  where $\hat{\boldsymbol{R}}_K = \sum_{i=K+1}^{p} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' = (\hat{r}_{ij})_{p \times p}$ is the principal orthogonal complement
- Keep the covariance matrix that is formed by the first K principal components and apply the thresholding procedure to the remaining covariance matrix $\hat{\boldsymbol{R}}_K$

$$\hat{\mathbf{R}}_K^{\mathcal{T}} = (\hat{r}_{ij}^{\mathcal{T}})_{p \times p}, \qquad \hat{r}_{ij}^{\mathcal{T}} = \begin{cases} \hat{r}_{ii}, & i = j, \\ s_{ij}(\hat{r}_{ij})\, I(|\hat{r}_{ij}| \geqslant \tau_{ij}), & i \neq j, \end{cases}$$

Sparsity on $\hat{\boldsymbol{R}}_K$:

$$\hat{\mathbf{R}}_K^{\mathcal{T}} = (\hat{r}_{ij}^{\mathcal{T}})_{p\times p}, \qquad \hat{r}_{ij}^{\mathcal{T}} = \begin{cases} \hat{r}_{ii}, & i = j, \\ s_{ij}(\hat{r}_{ij})\, I(|\hat{r}_{ij}| \geqslant \tau_{ij}), & i \neq j, \end{cases}$$

$s_{ij}(\cdot)$ is a generalized shrinkage function (Antoniadis & Fan, 2001)[4] and $\tau_{ij} > 0$ is an entry-dependent threshold. In particular:

$$s_{ij}(x) = x\boldsymbol{I}(|x| \geq \tau_{ij}) \text{ and } \tau_{ij} = \tau(\hat{r}_{ii}\hat{r}_{jj})^{1/2} \text{ for a given } \tau$$

Finally, the estimator of $\boldsymbol{\Sigma}$ is defined as

$$\hat{\boldsymbol{\Sigma}}_K = \sum_{i=1}^{K} \hat{\lambda}_i \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i' + \hat{\boldsymbol{R}}_K^{\mathcal{T}}$$

Some special cases

- when $\tau = 0$, the estimator is the sample covariance matrix
- when $\tau = 1$, the estimator becomes is based on the strict factor model
- when $K = 0$, the estimator is the same as the thresholding estimator of Bickel and Levina (2008)

When the number of factors K is unknown, it can be estimated from the data

[4] Regularized wavelet approximations. JASA

POET with unknown K

The proposed method allows a data-driven $\hat{K}$ to estimate the covariance matrices. The authors apply the well-know method in Bai and Ng (2002):

$$\hat{K} = \arg \min_{0 \leqslant K_1 \leqslant M} \log\left(\frac{1}{pT} \| \mathbf{Y} - T^{-1} \mathbf{Y} \hat{\mathbf{F}}_{K_1} \hat{\mathbf{F}}'_{K_1} \|_{\mathrm{F}}^2 \right) + K_1 \, g(T, p), \tag{2.14}$$

where M is a prescribed upper bound, $\hat{\mathbf{F}}_{K_1}$ is a $T \times K_1$ matrix whose columns are $\sqrt{T}$ times the eigenvectors corresponding to the K_1 largest eigenvalues of the $T \times T$ matrix $\mathbf{Y}'\mathbf{Y}$ and $g(T, p)$ is a penalty function of (p, T) such that $g(T, p) = o(1)$ and $\min\{p, T\} \, g(T, p) \to \infty$. Two examples suggested by Bai and Ng (2002), IC1 and IC2, are respectively

$$g(T, p) = \frac{p + T}{pT} \log\left(\frac{pT}{p + T} \right),$$

$$g(T, p) = \frac{p + T}{pT} \log(\min\{p, T\}).$$

Throughout the paper, we let $\hat{K}$ be the solution to problem (2.14) by using either IC1 or IC2.

Convergence of POET

One of the main result of the paper is the rate of convergence of the adaptive thresholding estimator

$$\|\hat{\Sigma}^{\mathrm{T}}_{u,\hat{K}} - \Sigma_u\| = O_p\left(m_p\left[\frac{1}{\sqrt{p}} + \sqrt{\left\{\frac{\log(p)}{T}\right\}}\right]^{1-q}\right),$$

where $q \in [0,1)$. $\hat{\Sigma}^{\mathrm{T}}_{u,\hat{K}}$ is the thresholded version of Σ_u, and $m_p = \max_{i \leqslant p} \sum_{j \leqslant p} I_{(\sigma_{u,ij} \neq 0)}$ when $q = 0$.

Choice of threshold

Threshold value: $\tau_{ij} = C\omega_T\sqrt{\hat{\theta}_{ij}}$

where $\hat{\theta}_{ij} = (1/T)\sum_{t=1}^{T}(\hat{u}_{it}\hat{u}_{jt} - \hat{\sigma}_{ij})^2$, $\omega_T = (1/\sqrt{p}) + \sqrt{\log(p)/T}$ and $C > 0$ is determined by the users.

We choose C in the range where $\lambda_{min}(\hat{\mathbf{\Sigma}}^{\mathcal{T}}_{u,\hat{K}}) > 0$

Define

$$C_{\min} = \inf[C > 0 : \lambda_{\min}\{\hat{\Sigma}^{\mathcal{T}}_{u,\hat{K}}(M)\} > 0, \quad \forall M > C]. \tag{4.1}$$

When C is sufficiently large, the estimator becomes diagonal, whereas its minimum eigenvalue must retain strict positivity. Thus, $C_{\min}$ is well defined and, for all $C > C_{\min}$, $\hat{\Sigma}^{\mathcal{T}}_{u,\hat{K}}(C)$ is positive definite under finite samples. We can obtain $C_{\min}$ by solving $\lambda_{\min}\{\hat{\Sigma}^{\mathcal{T}}_{u,\hat{K}}(C)\} = 0, C \neq 0$.

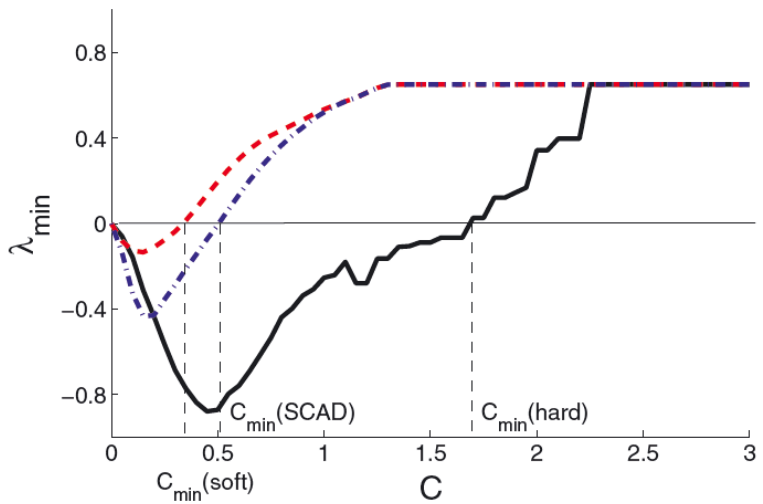

Fig. 1. Minimum eigenvalue of $\hat{\Sigma}^{\mathcal{T}}_{u\hat{K}}(C)$ as a function of C for three choices of thresholding rules (the plot is based on the simulated data set in Section 6.2): ——, hard thresholding; – – –, soft thresholding; ·–·–·, smoothly clipped absolute deviation

Experiments: robustness to the estimation of K

6.2. Simulation

For the simulation, we fix $T = 300$, and let p increase from 1 to 600. For each fixed p, we repeat the following steps $N = 200$ times, and record the means and the standard deviations of each respective norm.

Step 1: generate independently $\{\mathbf{b}_i\}_{i=1}^{p} \sim N_3(\mu_B, \Sigma_B)$, and set $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$.
Step 2: generate independently $\{\mathbf{u}_t\}_{t=1}^{T} \sim N_p(\mathbf{0}, \Sigma_u)$.
Step 3: generate $\{\mathbf{f}_t\}_{t=1}^{T}$ as a vector auto-regressive sequence of the form $\mathbf{f}_t = \mu + \Phi \mathbf{f}_{t-1} + \varepsilon_t$.
Step 4: calculate $\{\mathbf{y}_t\}_{t=1}^{T}$ from $\mathbf{y}_t = \mathbf{B}\mathbf{f}_t + \mathbf{u}_t$.
Step 5: set hard thresholding with threshold $0.5\sqrt{\hat{\theta}_{ij}}[\sqrt{\{\log(p)/T\}} + 1/\sqrt{p}]$. Estimate K by using IC1 of Bai and Ng (2002). Calculate covariance estimators by using POET. Calculate the sample covariance matrix $\hat{\Sigma}_{\text{sam}}$.

To assess the robustness, they calculate $\hat{\Sigma}_{u,K}^{\mathcal{T}}$ for $K = 1, 2, \ldots, 10$ and calculate different norms:

$$\|\hat{\Sigma}_{u,K}^{\mathcal{T}} - \Sigma_u\|, \|(\hat{\Sigma}_{u,K}^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\|, \|\hat{\Sigma}_K^{-1} - \Sigma^{-1}\| \text{ and } \|\hat{\Sigma}_K - \Sigma\|_\Sigma$$

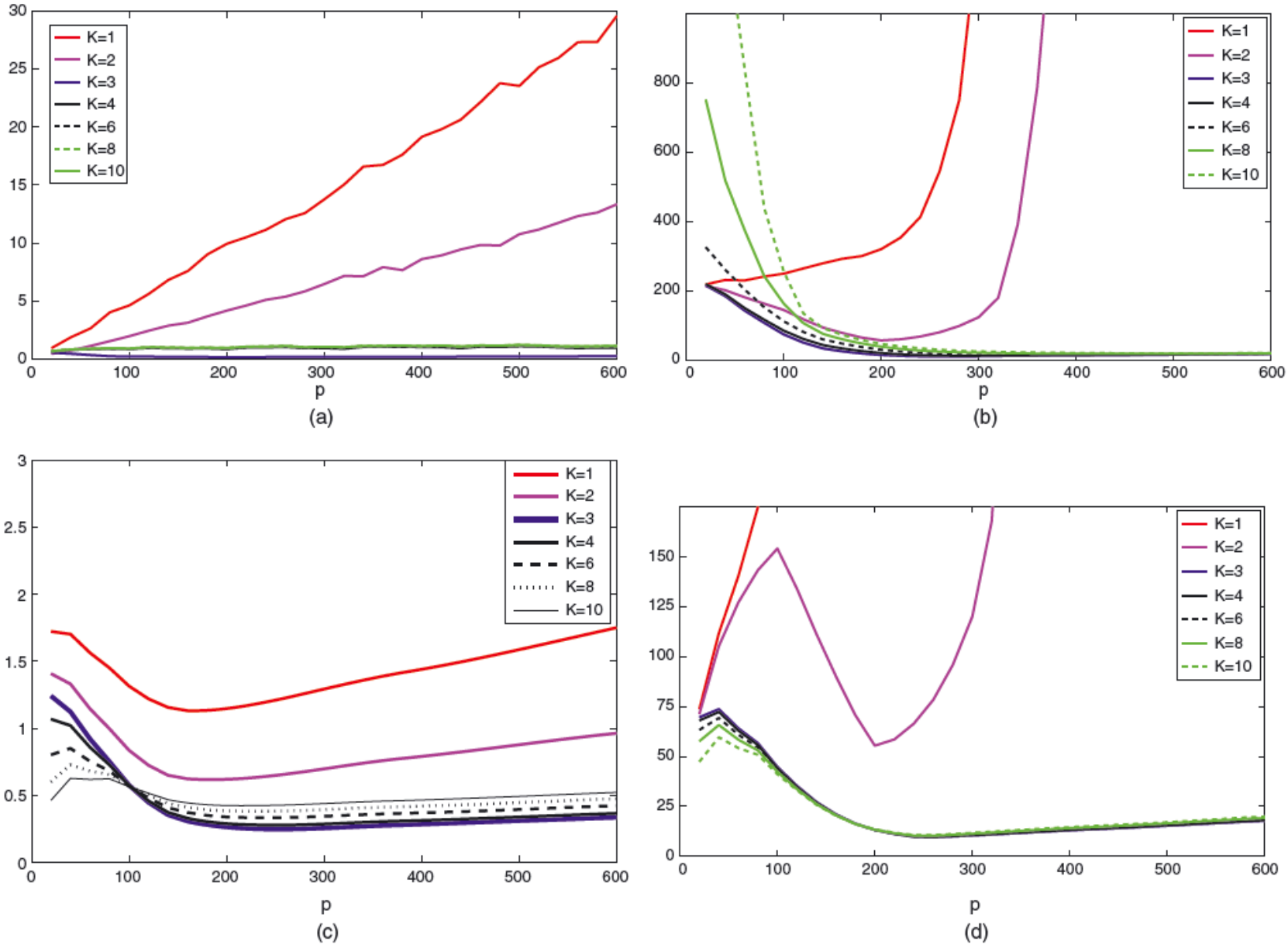

Fig. 6. Robustness of K as p increases for various choices of K (design 1, $T = 300$): (a) $\|\hat{\Sigma}_{u,K}^{\mathcal{T}} - \Sigma_u\|$; (b) $\|(\hat{\Sigma}_{u,K}^{\mathcal{T}})^{-1} - \Sigma_u^{-1}\|$; (c) $\|\hat{\Sigma}_K - \Sigma\|_u$; (d) $\|\hat{\Sigma}_K - \Sigma^{-1}\|$

Experiments: comparisons with other methods

They compare POET with related methods that address low rank + sparse covariance estimation: (i) LOREC: low rank and sparse covariance estimator proposed by Luo (2011); (ii) SFM: strict factor model by Fan et (2008); (iii) DUAL: the dual method by Lin et al (2009); and (iv) SVT: the singular value thresholding method of Cai et al (2008)

Table 4. Method comparison under spectral norm for $T = 100$†

Method	$\hat{\Sigma}_u$	$\hat{\Sigma}_u^{-1}$	*RelE*	$\hat{\Sigma}^{-1}$	$\hat{\Sigma}$
p=100					
POET	1.624	1.336	2.080	1.309	29.107
LOREC	2.274	1.880	2.564	1.511	32.365
SFM	2.084	2.039	2.707	2.022	34.949
Dual	2.306	5.654	2.707	4.674	29.000
SVT	2.59	13.64	2.806	103.1	29.670
p=200					
POET	1.641	1.358	3.295	1.346	58.769
LOREC	2.179	1.767	3.874	1.543	62.731
SFM	2.098	2.071	3.758	2.065	60.905
Dual	2.41	6.554	4.541	5.813	56.264
SVT	2.930	362.5	4.680	47.21	63.670
p=300					
POET	1.662	1.394	4.337	1.395	65.392
LOREC	2.364	1.635	4.909	1.742	91.618
SFM	2.091	2.064	4.874	2.061	88.852
Dual	2.475	2.602	6.190	2.234	74.059
SVT	2.681	$> 10^3$	6.247	$> 10^3$	80.954

†RelE represents the relative error $\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbf{I}_p\|$.

Real data: Sparsity of idiosyncratic errors

Data from the Center for Research in Security Prices database, for $p = 50$ stocks and daily returns (Jan 1st, 2010 - Dec 31st, 2010) $T = 252$. Stock chosen from five different industry sectors (consumer goods, financial, healthcare, services, utilities) with 10 stocks from each sectors

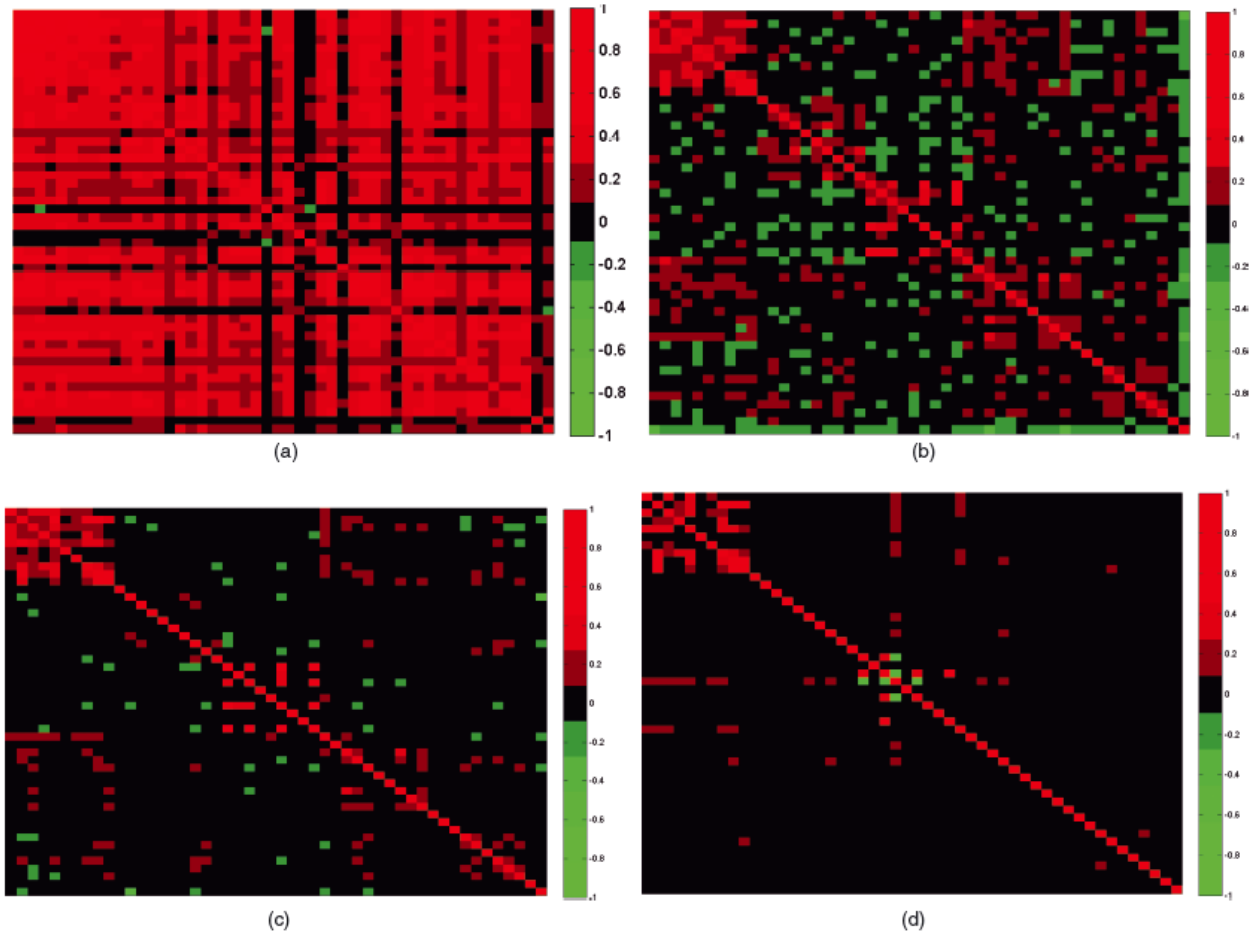

Fig. 9. Heat map of the thresholded error correlation matrix for number of factors (a) $K = 0$, (b) $K = 1$, (c) $K = 2$ and (d) $K = 3$

Discussion

The authors study the problem of estimating high dimensional covariance matrix with *conditional sparsity*

Realizing that an *unconditional sparsity* assumption is inappropriate in many applications, they introduce a latent factor model (with conditional sparsity) and propose the POET method

The method expands considerably the scope of the strict factor model which assumes independent idiosyncratic noise and is too restrictive in practice

By assuming a sparse error covariance matrix we allow for the presence of the cross-sectional correlation even after taking out the common factors

The method has wide applicability in statistical genomics to estimate the covariance and its network model

R package: http://cran.r-project.org/web/packages/POET/index.html