Electronic Journal of Statistics
Vol. 10 (2016) 60–66
ISSN: 1935-7524
DOI: 10.1214/15-EJS1018

Discussion of "Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation"*

Hui Zou

School of Statistics
University of Minnesota
e-mail: hzou@stat.umn.edu

Abstract: Professors Cai, Ren and Zhou ought to be congratulated for writing such a wonderful expository paper on optimal estimation of high-dimensional covariance and precision matrices. Nearly all optimality results on large matrix estimation were established by the authors (and their co-authors). Thus, they are the most appropriate team to write this much needed review article. My discussion contains three sections.

Received March 2015.

Contents

1. SURE information criteria

Cai, Zhang and Zhou (2010) developed the first minimax optimality result for estimating bandable covariance matrices. They derived the minimax rates of convergence under the matrix ℓ_2, ℓ_1 and Frobenius norms where the parameter space is defined as

$$\mathcal{F}_\alpha = \{\mathbf{\Sigma} : \max_j \sum_i \{|\sigma_{ij}| : |i-j| > k\} \leq Mk^{-\alpha} \, \forall k, \text{and} \, \lambda_{\max}(\Sigma) \leq M_0\}. \quad (1)$$

A tapering covariance estimator was constructed to achieve the minimax bound and its tapering bandwidth differs under difference matrix norms. The optimal tapering bandwidth is $n^{\frac{1}{2\alpha+2}}$ under Frobenius norm and $n^{\frac{1}{2\alpha+1}}$ under the ℓ_2 norm.

*Main article 10.1214/15-EJS1081.

The ℓ_2 norm is a preferred metric if the estimated covariance matrix is used in another estimation problem. Frobenius norm is a good metric for checking the goodness of fit. Frobenius norm can serve as a very good tuning metric for selecting a covariance matrix estimator that can perform well under the ℓ_2 norm. There are at least two reasons for such a practice. First, an optimal covariance matrix estimator under Frobenius norm should also perform very well (although may not be optimal) under the ℓ_2 norm. For example, under Frobenius norm the minimax rate optimal tapering estimator uses a bandwidth $n^{\frac{1}{2\alpha+2}}$ and its ℓ_2 risk is bounded by $O(n^{-\frac{\alpha}{\alpha+1}} + \frac{\log(p)}{n})$. We can compare it to the minimax optimal rate under the ℓ_2 norm: $O(n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log(p)}{n})$. The difference is very small for large α. Let $\log(p) = n^\gamma$ for some $0 < \gamma < 1$, then for $\frac{1}{\alpha+1} \ll \gamma$, both bounds reduce to $O(n^{\gamma-1})$. Second, the Frobenius norm is smoother than the ℓ_2 norm which has an important practical consequence. It has been empirically demonstrated in Yi and Zou (2013) that when using cross-validation to select the tapering bandwidth, the chosen bandwidth under the ℓ_2 norm is highly variable while the chosen bandwidth under Frobenius norm is more stable.

Li and Zou (2014) extended Stein's unbiased risk estimation (SURE) theory to bandable covariance matrices estimation and proved its asymptotical optimality under the Frobenius norm. To provide a unified treatment of the banding estimator (Wu and Pourahmadi, 2003; Bickel and Levina, 2008) and the tapering estimator (Cai, Zhang and Zhou, 2010), we consider the generalized tapering estimator of the covariance matrix:

$$\widehat{\mathbf{\Sigma}}^{(\tau)} = \left(\hat{\sigma}_{ij}^{(\tau)}\right)_{1\le i,j\le p} = \left(\omega_{ij}^{(\tau)}\tilde{\sigma}_{ij}\right)_{1\le i,j\le p} \tag{2}$$

where $\widetilde{\mathbf{\Sigma}}$ is the MLE of $\mathbf{\Sigma}$ and the generic tapering weights $(\omega_{ij}^{(\tau)})_{1\le i,j\le p}$ should satisfy

(i) $\omega_{ij}^{(\tau)} = 1$ for $|i-j| \le \lfloor \frac{\tau}{2} \rfloor$,
(ii) $\omega_{ij}^{(\tau)} = 0$ for $|i-j| \ge \tau$,
(iii) $0 \le \omega_{ij}^{(\tau)} \le 1$ for $\lfloor \frac{\tau}{2} \rfloor < |i-j| < \tau$.

For any generalized tapering estimator $\widehat{\Sigma}^{(\tau)}$, its Frobenius risk is $R(\tau) = \mathbb{E}(\|\hat{\Sigma}^{(\tau)} - \Sigma\|_F^2)$. It is shown that (Yi and Zou, 2013; Li and Zou, 2014)

$$R(\tau) = \mathbb{E}(\mathrm{SURE}(\tau)) \tag{3}$$

$$\begin{aligned}\mathrm{SURE}(\tau) = &\sum_{1\le i,j\le p} (\frac{n}{n-1} - \omega_{ij}^{(\tau)})^2 \tilde{\sigma}_{ij}^2 \\ &+ \sum_{1\le i,j\le p} (2\omega_{ij}^{(\tau)} - \frac{n}{n-1})(a_n\tilde{\sigma}_{ij}^2 + b_n\tilde{\sigma}_{ii}\tilde{\sigma}_{jj})\end{aligned} \tag{4}$$

with

$$a_n = \frac{n(n-3)}{(n-1)(n-2)(n+1)} \text{ and } b_n = \frac{n}{(n+1)(n-2)}. \tag{5}$$

We pick the best tapering estimator from the list of all tapering estimators by minimizing SURE(τ) as a function of τ. The chosen tapering estimator is $\widehat{\mathbf{\Sigma}}^{(\hat{\tau}_n)}$ where

$$\hat{\tau}_n = \arg\min_{\tau} \text{SURE}(\tau). \tag{6}$$

We call this method SURE tuning. The following theorem shows that SURE tuning achieves the minimax optimal rate of convergence under Frobenius norm.

Theorem 1 (Li and Zou, 2014)**.** *Assume* $n \leq p$ *and* $\log p = o(n)$*, then* $\sup_{\Sigma \in \mathcal{F}_\alpha} \mathbb{E}\|\widehat{\mathbf{\Sigma}}^{(\hat{\tau}_n)} - \mathbf{\Sigma}\|_F^2 \asymp pn^{-(2\alpha+1)/2(\alpha+1)}$.

If the true covariance matrix is exactly banded with a bandwidth k_0, then it is shown that SURE tuning almost surely selects a larger bandwidth. See Theorem 6 in Li and Zou (2014). In order to achieve consistent selection, Li and Zou (2014) further modified SURE criterion as follows:

$$\begin{aligned} \text{SURE}_{\log n}(\tau) \quad = \quad & \sum_{1 \leq i,j \leq p} (\frac{n}{n-1} - \omega_{ij}^{(\tau)})^2 \tilde{\sigma}_{ij}^2 \\ & + \sum_{1 \leq i,j \leq p} (\log(n)\omega_{ij}^{(\tau)} - \frac{n}{n-1})(a_n \tilde{\sigma}_{ij}^2 + b_n \tilde{\sigma}_{ii}\tilde{\sigma}_{jj}). \end{aligned} \tag{7}$$

Note that we replace the constant 2 with $\log(n)$ to go from SURE to $\text{SURE}_{\log n}$. Since the true covariance matrix is banded, we select a banding estimator from the list of all banding estimators. So $\omega_{ij}^{(\tau)} = I(|i-j| < \tau)$.

Theorem 2 (Li and Zou, 2014)**.** *Let the true covariance matrix* $\mathbf{\Sigma}$ *be a banded matrix with bandwidth* k_0 *such that* $\sigma_{ij} = 0$ *if* $|i-j| \geq k_0$ *and* $\min_{|i-j| \leq k_0 - 1} \sigma_{ij}^2 \gg \log n/n$*, where* k_0 *is a constant doesn't depend on* n*. If* $n \leq p_n$ *and* $\log p_n = o(n)$*,* $k_0 = \arg\min_\tau SURE_{\log n}(\tau)$ *almost surely.*

Theorem 1 and Theorem 2 suggest that SURE and $\text{SURE}_{\log n}$ can be regarded as AIC and BIC for estimating large bandable covariance matrices, respectively.

2. Sparse recovery in precision matrix estimation

Sparse precision matrix estimation has received a lot of attention recently due to its immediate application to graphical models. A sparse precision matrix of a normal distribution can be easily turned into a Gaussian graphical model where the nonzero entries in the precision matrix correspond to the edges in the graph. So far, we have seen three approaches for constructing a sparse precision matrix estimator. The first one is by using a sparse penalized likelihood (Yuan and Lin, 2007; Ravikumar et al., 2008; Rothman et al., 2008; Friedman et al., 2008). The second is the so-called neighborhood regression estimation where the method tries to estimate each column (or row) of the precision matrix one by one and then combines them into a matrix. See the neighborhood lasso regression estimator (Meinshausen and Bühlmann, 2006), the neighborhood Dantzig selector (Yuan, 2010) and the neighborhood scaled lasso estimator (Sun and Zhang,

2012). The third approach uses a direct constrained minimization criterion, see CLIME by Cai, Liu and Luo (2011). Each approach has its own distinct merits and drawbacks. The penalized loss approach always gives a positive definite matrix, while the other two do not even guarantee the estimator is symmetric and some postprocessing is needed. The neighborhood approach is computationally very friendly: there are many software for doing sparse penalized regression. The computation of CLIME can be done in parallel as well and CLIME enjoys a nice rate of convergence without assuming any difficult structure assumption, such as the irrepresentable conditions required in the lasso penalized likelihood and neighborhood lasso regression.

Fan, Xue and Zou (2014) proposed a hybrid method to obtain a sparse precision matrix estimator that is also positive definite and enjoys strong oracle property without requiring the irrepresentable condition. The hybrid estimator uses the CLIME estimator as its initial value and then performs the weighted ℓ_1 penalized likelihood estimation twice.

The hybrid precision matrix estimator by Fan, Xue and Zou (2014)

1. Initialize $\widehat{\boldsymbol{\Theta}}^{(0)} = \widehat{\boldsymbol{\Theta}}^{\text{clime}}$ where
$$\widehat{\boldsymbol{\Theta}}^{\text{clime}} = \arg\min_{\boldsymbol{\Theta}} \|\boldsymbol{\Theta}\|_1 \quad \text{subject to } \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta} - \boldsymbol{I}\|_{\max} \le \lambda_{clime}. \tag{8}$$
2. Compute $\hat{w}_{ij}^{(1)} = P_\lambda'(|\hat{\theta}_{ij}^{(0)}|)$ and solve $\widehat{\boldsymbol{\Theta}}^{(1)}$ from
$$\widehat{\boldsymbol{\Theta}}^{(1)} = \arg\min_{\boldsymbol{\Theta}\succ 0} \left\{ -\log\det(\boldsymbol{\Theta}) + \langle \boldsymbol{\Theta}, \widehat{\boldsymbol{\Sigma}}_n \rangle + \sum_{(i,j),i\neq j} \lambda \hat{w}_{ij}^{(1)} |\theta_{ij}| \right\}. \tag{9}$$
3. Compute $\hat{w}_{ij}^{(2)} = P_\lambda'(|\hat{\theta}_{ij}^{(1)}|)$ and solve $\widehat{\boldsymbol{\Theta}}^{(2)}$ from
$$\widehat{\boldsymbol{\Theta}}^{(2)} = \arg\min_{\boldsymbol{\Theta}\succ 0} \left\{ -\log\det(\boldsymbol{\Theta}) + \langle \boldsymbol{\Theta}, \widehat{\boldsymbol{\Sigma}}_n \rangle + \sum_{(i,j),i\neq j} \lambda \hat{w}_{ij}^{(2)} |\theta_{ij}| \right\}. \tag{10}$$
4. Report $\widehat{\boldsymbol{\Theta}}(\lambda_{clime}, \lambda) = \widehat{\boldsymbol{\Theta}}^{(2)}$ as the final estimator.

We need some notation to present the theorem justifying the above hybrid estimator. Let the true precision matrix $\boldsymbol{\Theta}^\star = (\theta_{jk}^\star)_{q\times q}$ with the support set $\mathcal{A} = \{(j,k) : \theta_{jk}^\star \neq 0\}$. Write $L = \|\boldsymbol{\Theta}^\star\|_1$, $\|\boldsymbol{\Theta}_{\mathcal{A}}^\star\|_{\min} = \min\{|\theta_{jk}^\star| : i,j \in \mathcal{A}\}$, $s = \#\{(j,k) : j \le k, \theta_{jk}^\star \neq 0\}$ and $d = \max_j \#\{k : \theta_{jk}^\star \neq 0\}$. Let $\boldsymbol{H}^\star = \boldsymbol{\Sigma}^\star \bigotimes \boldsymbol{\Sigma}^\star$ and define

$$K_1 = \|\boldsymbol{\Sigma}^\star\|_{\ell_\infty}, \quad K_2 = \|(\boldsymbol{H}_{\mathcal{A}\mathcal{A}}^\star)^{-1}\|_{\ell_\infty}, \quad \text{and} \quad K_3 = \|\boldsymbol{H}_{\mathcal{A}^c\mathcal{A}}^\star (\boldsymbol{H}_{\mathcal{A}\mathcal{A}}^\star)^{-1}\|_{\ell_\infty}.$$

The theoretical oracle estimator is defined as

$$\widehat{\boldsymbol{\Theta}}^{oracle} = \arg\min_{\boldsymbol{\Theta}\succ 0:\ \boldsymbol{\Theta}_{\mathcal{A}^c}=0} \left\{ -\log\det(\boldsymbol{\Theta}) + \langle \boldsymbol{\Theta}, \widehat{\boldsymbol{\Sigma}}_n \rangle \right\}. \tag{11}$$

Theorem 3 (Fan, Xue and Zou, 2014)**.** *If* $\frac{\|\mathbf{\Theta}^\star_{\mathcal{A}}\|_{\min}}{a+1} > \lambda > \frac{4L}{a_0}\lambda_{clime}$*, then* $\widehat{\mathbf{\Theta}}(\lambda_{clime}, \lambda)$ *is equal to the oracle estimator* $\widehat{\mathbf{\Theta}}^{oracle}$ *with probability at least* $1 - \delta_1 - \delta_2 - C_0 p \exp(-\frac{c_0 n}{L^2}\lambda^2_{clime})$*, where*

$$\begin{aligned}
\delta_1 &= C_0 s \cdot \exp\big(-\frac{c_0}{4} n \cdot \min\big\{\frac{a_1^2\lambda^2}{(2K_3+1)^2}, \frac{1}{9K_1^2K_2^2d^2}, \frac{1}{9K_1^6K_2^4d^2}\big\}\big) \\
&\quad + C_0(p-s)\cdot\exp\big(-\frac{c_0 a_1^2}{4} n\lambda^2\big), \\
\delta_2 &= C_0 s \cdot \exp\big(-\frac{c_0 n}{4K_2^2} \cdot \min\big\{\frac{1}{9K_1^2d^2}, \frac{1}{9K_1^6K_2^2d^2}, (\|\mathbf{\Theta}^\star_{\mathcal{A}}\|_{\min} - a\lambda)^2\big\}\big),
\end{aligned}$$

and C_0, c_0 *are constants.*

3. Semiparametric Gaussian copulas

In many covariance/precision matrices estimation problems the multivariate normal distribution assumption about the data is not needed and the rates of convergence can be established by assuming certain tail probability bounds. The authors have explained this fact clearly in this review article. On the other hand, it is often desirable to have normality in practice. For example, normality is a key link that bridges a sparse precision matrix and a Gaussian graphical model. Even from theoretical viewpoint, normality is also helpful, because it yields faster rates of convergence than polynomial tail bounds. The difficulty here is that observed data often violate normality: the distribution can be heavily skewed, multimodal or have heavy tails.

Semiparametric Gaussian copulas offer a nice compromise between the reality and normality by assuming the data after univariate monotone transformations will follow a multivariate normal distribution.

> **The Semiparametric Gaussian Copula Model:** $(X_1, \ldots, X_p)$ *follows a p-dimensional semiparametric Gaussian copula, if there exists a vector of unknown univariate monotone increasing transformations, denoted by* $(f_1, \ldots, f_p)$*, such that the transformed random vector follows a multivariate normal distribution with mean 0 and covariance* $\mathbf{\Sigma}$*:*
>
> $$(f_1(X_1), \ldots, f_p(X_p)) \sim N_p(0, \mathbf{\Sigma}), \tag{12}$$
>
> *where without loss of generality the diagonals of* $\mathbf{\Sigma}$ *are equal to 1.*

Note that the marginal normality is always achieved by transformations, so the key assumption is that those marginally normal-transformed variables are jointly normal, which is the parametric component of the semiparametric Gaussian copula model.

It is interesting to see that the conditional dependence/independence relations among X_js can be seen from the precision matrix $\mathbf{\Sigma}^{-1}$. Hence, the graphical model problem is still equivalent to the problem of estimating a sparse

precision matrix under the semiparametric Gaussian copula model. If we knew the transformation functions, then we could work on the transformed data and apply methods such as the penalized likelihood and CLIME to estimate $\mathbf{\Sigma}^{-1}$. Xue and Zou (2012) showed a systematic rank-based inference approach to estimating $\mathbf{\Sigma}^{-1}$ without involving any estimators of these p many univariate nonparametric monotone functions. Define the rank correlation matrix as follows:

$$\hat{\boldsymbol{R}}^s = (\hat{r}^s_{ij})_{1\le i,j\le p}. \tag{13}$$

where

$$\hat{r}^s_{ij} = 2\sin(\frac{\pi}{6}\hat{r}_{ij}). \tag{14}$$

and $\hat{r}_{ij}$ is the Spearman's correlation between variables i and j. One can feed $\hat{\boldsymbol{R}}^s$ to a regular sparse precision matrix method to obtain the corresponding rank-based counterpart. For example, the rank-based graphical lasso estimator is defined as

$$\widehat{\mathbf{\Theta}}^s_g = \arg\min_{\mathbf{\Theta}\succ 0}\left\{-\log\det(\mathbf{\Theta}) + \langle\hat{\boldsymbol{R}}^s, \mathbf{\Theta}\rangle + \lambda\sum_{i\neq j}|\theta_{ij}|\right\}. \tag{15}$$

The rank-based CLIME is defined as

$$\widehat{\mathbf{\Theta}}^s_c = \arg\min_{\mathbf{\Theta}}\|\mathbf{\Theta}\|_1 \quad \text{subject to} \quad \|\hat{\boldsymbol{R}}^s\mathbf{\Theta} - \boldsymbol{I}\|_{\max}\le\lambda. \tag{16}$$

Xue and Zou (2012) showed that the rank-based estimators work as well as the corresponding "oracle" estimators by using the transformed "oracle" data.

Rank-based estimation of $\mathbf{\Sigma}$ under bandable or sparse structure assumptions has been considered in Xue and Zou (2014) and Xue and Zou (2013).

Mai and Zou (2015) used the semiparametric copula model to develop a sparse semiparametric discriminant analysis classifier where the estimation of the monotone transformation functions are required. Mai and Zou (2015) derived a uniformly consistent estimator of $(f_1,\ldots,f_p)$ under the assumption that $\log(p) = O(n^\gamma)$ with $\gamma < \frac{1}{3}$. It would be interesting to further raise the upper bound on γ to 1.

Acknowledgements

I thank Professor George Michailidis for inviting me to contribute this discussion.

References

Fan, J., Xue, L. and Zou, H. (2014). Strong Oracle Optimality of Folded Concave Penalized Estimation. *Annals of Statistics* **42** 819–849. MR3210988

Li, D. and Zou, H. (2014). SURE Information Criteria for Large Covariance Matrix Estimation and Their Asymptotic Properties. *arXiv:1406.6514*.

Mai, Q. and Zou, H. (2015). Sparse Semiparametric Discriminant Analysis. *Journal of Multivariate Analysis* **135** 175–188. MR3306434

Xue, L. and Zou, H. (2012). Regularized Rank-based Estimation of High-dimensional Nonparanormal Graphical Models. *Annals of Statistics* **40** 2541–2571. MR3097612

Xue, L. and Zou, H. (2013). Optimal Estimation of Sparse Correlation Matrices of Semiparametric Gaussian Copulas. *Statistics and Its Interface* **7** 201–209. MR3199378

Xue, L. and Zou, H. (2014). Rank-based Tapering Estimation of Bandable Correlation Matrices. *Statistica Sinica* **24** 83–100. MR3183675

Yi, F. and Zou, H. (2013). SURE-tuned Tapering Estimation of Large Covariance Matrices. *Computational Statisticsa and Data Analysis* **58** 339–351. MR2997947