Electronic Journal of Statistics
Vol. 10 (2016) 81–89
ISSN: 1935-7524
DOI: 10.1214/15-EJS1081REJ

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

T. Tony Cai

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: tcai@wharton.upenn.edu

Zhao Ren

Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA
e-mail: zren@pitt.edu

and

Harrison H. Zhou

Department of Statistics, Yale University, New Haven, CT 06511, USA
e-mail: huibin.zhou@yale.edu

Received January 2016.

1. Introduction

We are deeply grateful to the discussants for providing constructive and stimulating comments and suggestions. Our paper gives a survey of recent optimality and adaptivity results on estimating various families of structured covariance and precision matrices in the high-dimensional setting, with a focus on understanding the intrinsic difficulty of the problems. To achieve this goal, we present main results under relative simple and unified assumptions, and hence do not address some practical issues. Several such questions are raised by the discussants, including robustness to outliers, Gaussian assumption, estimation with missing data, estimation with time-dependent observations, alternative bandwidth selection, and hybrid procedures for estimating precision matrices. In this rejoinder we comment on these important points and remark on some challenges that lie ahead.

*Main article 10.1214/15-EJS1081.

2. Robust estimation

Professor Zou as well as Professors Balasubramanian and Yuan raise questions on the Gaussian or sub-Gaussian assumption and discuss possible robust methods when this assumption fails to hold. Robust estimation is an important goal and deserves a careful investigation in the problems studied in the present paper.

2.1. Semiparametric Gaussian copula model

Professor Zou suggests to consider the semiparametric Gaussian copula model, in which the data after certain univariate monotone transformations follow a multivariate Gaussian distribution. Specifically, n copies of random vector $X = (X_1, \ldots, X_p)'$ are observed, where $f(X) = (f_1(X_1), \ldots, f_p(X_p))' \sim N(0, \Sigma)$ and $\{f_i\}_{i=1}^p$ are unknown strictly increasing functions. Due to identifiability issue, the goal is to estimate structured correlation matrix or its inverse.

There has been much recent attention on various high-dimensional statistical problems under the semiparametric Gaussian copula model. In particular, several estimation problems surveyed in the current paper have been considered. Almost all these methods are based on the rank-based correlation matrix estimators: $\hat{R} = (\hat{r}_{ij}) = \sin\left(\frac{\pi}{2}\hat{\tau}_{ij}\right)$ or $\hat{R} = (\hat{r}_{ij}) = 2\sin\left(\frac{\pi}{6}\hat{\rho}_{ij}\right)$, where $\hat{\tau}_{ij}$ and $\hat{\rho}_{ij}$ are Kendall's tau and Spearman's rho estimators respectively. Two fundamental concentration inequalities have been estalished: (i) entrywise deviation bound $|\hat{r}_{ij} - \sigma_{ij}|$, (Xue and Zou (2012) and Liu et al. (2012)) and (ii) submatrix deviation bound under the spectral norm loss $||(\hat{R} - \Sigma)_{S,S}||$ with some index set S (Mitra and Zhang (2014), Wegkamp and Zhao (2016) and Han and Liu (2013)). In particular, both bounds enjoy the same properties as those obtained by using the unobservable sample covariance matrix of $f(X^{(i)})$.

We briefly discuss some results on optimal estimation of covariance and precision matrices based on $\hat{R}$. Xue and Zou (2013) studied optimal estimation of sparse covariance matrices over the class $\mathcal{H}(c_{n,p})$, noticing that the analysis only relies on bound (i). When replacing the sample covariance matrix by $\hat{R}$ in the CLIME estimator and graphical Lasso estimator in (27) respectively, Xue and Zou (2012) and Liu et al. (2012) obtained similar theoretical results for estimating sparse precision matrices using the entrywise deviation bound (i). With the help of the concentration bound (ii), Mitra and Zhang (2014) established optimal rate over the class of bandable matrices $\mathcal{F}_\alpha(M_0, M)$ under the spectral norm loss (also see Xue and Zou (2014)). Recently, Barber and Kolar (2015) took a similar approach to estimate individual entries of sparse precision matrices over $\mathcal{HP}(c_{n,p}, M)$ using both bounds (i) and (ii). It is worth mentioning that the asymptotic efficiency result cannot be obtained due to the unknown f.

Without too much additional difficulty, the semiparametric Gaussian copula model can be further extended to semiparametric elliptical copula model (Liu et al. (2012)) in which $f(X)$ follows an elliptical distribution with scatter matrix Σ.

2.2. Robustness to contamination

Professors Balasubramanian and Yuan raise an interesting question on robust covariance matrix estimation against contamination or outliers among the data. This is indeed an important question as the statistical performance of the optimal estimators can be completely compromised with only one arbitrary outlier. This is also reflected in the numerical example provided by Professors Balasubramanian and Yuan. In their commentary, to this end, a simple median covariance matrix $\hat{\Sigma}^{\text{med}} = \arg\min_{A\in\mathbb{R}^{p\times p}} \sum_{i=1}^{n} \|X^{(i)}X^{(i)\prime} - A\|_F$ is proposed to replace the role of the usual sample covariance matrix in the covariance estimation procedures, motivated by the geometric median of a vector. The experiment shows a very promising result on a bandable covariance estimation example.

Recently Chen et al. (2015) investigated this issue under Huber's ϵ-contamination model $(1-\epsilon)P_\Sigma + \epsilon Q$ in the minimax decision framework. Under this model, P_Σ is the target distribution with the covariance matrix Σ and Q is some arbitrary contamination distribution. The goal is to estimate Σ optimally in some structured parameter spaces robust to the contamination. Chen et al. (2015) introduced a new concept called matrix depth motivated by Tukey's depth for vector estimation (Tukey, 1975) and proposed a robust covariance matrix estimator by maximizing the empirical depth function. The proposed estimator was shown to be minimax rate optimal under Huber's framework for bandable class $\mathcal{F}_\alpha(M_0, M)$ and sparse spiked covariance class $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})$. Interestingly, the optimal rates for classes $\mathcal{F}_\alpha(M_0, M)$ and $\mathcal{J}(c_{n,p}, r_{n,p}, \lambda_{n,p})$ are $\min\{\frac{\log p}{n} + n^{-\frac{2\alpha}{2\alpha+1}}, \frac{p}{n}\} \vee \epsilon^2$ and $\frac{c_{n,p}}{n}\log\frac{ep}{c_{n,p}} \vee \epsilon^2$ respectively.

Compared with the depth-based optimal estimator, median covariance matrix is obtained using the Frobenius norm. It is unclear to what extent this median covariance estimator leads to optimal rates with the presence of contamination under the spectral norm loss. Much work is still needed to gain fundamental understanding of robust estimation under ϵ-contamination model or other reasonable models.

3. Estimation with missing data

Motivated by the matrix completion problems as well as many high-dimensional applications, Professor Tsybakov raises an interesting question on estimating covariance and precision matrices in the presence of missing data. Indeed, in the high-dimensional setting, missing data problem arises in various applications such as genetic studies, climate research and cosmology.

Covariance and precision matrix estimation with incomplete data has been investigated in a few recent papers Lounici (2013, 2014); Kolar and Xing (2012); Loh and Wainwright (2012); Cai and Zhang (2015). Lounici (2013) and Lounici (2014) considered the missing data problem in the setting where the sample $\{Y^{(1)}, \ldots, Y^{(n)}\}$ is observed with $Y_j^{(i)} = m_j^{(i)} X_j^{(i)}$, where $\{X^{(1)}, \ldots, X^{(n)}\}$ are i.i.d. copies of a random vector X, and $m_j^{(i)}$ $(1 \le i \le n, 1 \le j \le p)$ are i.i.d. Bernoulli variables with parameter $\delta \in (0, 1]$ and independent of other

variables. The case $\delta = 1$ corresponds to the standard setting of fully observed data. Under such a setting, the sample covariance matrix $\hat{\Sigma}_Y = \mathbf{Y}'\mathbf{Y}/n$ with $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(n)})'$ is no longer an unbiased estimator of the target covariance $\Sigma = \mathrm{Cov}(X)$ but a simple scaling procedure leads to the following unbiased estimator

$$\hat{\Sigma}^{(\delta)} = (\hat{\sigma}_{ij}^{(\delta)}) := \left(\delta^{-1} - \delta^{-2}\right) \mathrm{diag}(\hat{\Sigma}_Y) + \delta^{-2}\hat{\Sigma}_Y. \tag{1}$$

Naturally, one can replace the sample covariance matrix by $\hat{\Sigma}^{(\delta)}$ in most procedures designed for fully observed data and obtain the corresponding final estimators of structured covariance or precision matrices. Indeed, Lounici (2013) applied this strategy to solve a sparse PCA problem while Lounici (2014) is focused on covariance matrix estimation when Σ is of low rank or can be approximated by a low rank matrix. Kolar and Xing (2012) and Loh and Wainwright (2012) also used $\hat{\Sigma}^{(\delta)}$ (or its submatrices in a regression approach) to solve sparse precision matrix estimation problems under the spectral norm loss for the class of precision matrices similar to $\mathcal{HP}(c_{n,p}, M)$.

The technical analysis of the new estimators using unbiased pivotal estimator $\hat{\Sigma}^{(\delta)}$ heavily relies on a careful study of two fundamental concentration bounds. They are (i) entrywise deviation bound $|\hat{\sigma}_{ij}^{(\delta)} - \sigma_{ij}|$, and (ii) submatrix deviation bound under the spectral norm loss $||(\hat{\Sigma}^{(\delta)} - \Sigma)_{S,S}||S$ with some index set S (see, for instance, Lounici (2013), Kolar and Xing (2012), Loh and Wainwright (2012)). It turns out that the main difference between $\hat{\Sigma}^{(\delta)}$ and the sample covariance matrix $\hat{\Sigma}$ without missing data in terms of those two concentration bounds is that the role of sample size n for $\hat{\Sigma}$ is replaced by $n\delta^2$ for $\hat{\Sigma}^{(\delta)}$. This explicit dependency on δ is also implied by (1) since we only have approximately $n\delta^2$ samples to estimate each off-diagonal entry in Σ. Consequently, it is not surprising to see or expect the same dependency of the minimax optimal rates of convergence on δ for various structured covariance and precision matrices such as bandable covariance matrices $\mathcal{F}_\alpha(M_0, M)$ and sparse covariant matrices $\mathcal{H}(c_{n,p})$.

Recently, Cai and Zhang (2015) considered optimal estimation of covariance matrices under a general missing completely at random (MCR) model. Specifically, the Bernoulli variables $m_j^{(i)}$ $(1 \le i \le n, 1 \le j \le p)$ were only assumed to be independent of $\{X^{(1)}, \ldots, X^{(n)}\}$, but otherwise arbitrary. Minimax rates of convergence for estimating bandable covariance matrices in $\mathcal{F}_\alpha(M_0, M)$ and sparse covariant matrices in $\mathcal{H}(c_{n,p})$ are established under the spectral norm loss. Cai and Zhang (2015) introduced the so-called extended sample mean and extended sample covariance matrix in the missing data setting and proposed adaptive rate-optimal covariance matrix estimators based on the these quantities.

While these simple missing data settings are relatively well understood, much work is still needed to gain fundamental understanding of optimal covariance and precision matrix estimation under other important missing data scenarios.

4. Estimation with time-dependent observations

Professors Paul and Wang point out an important setting when observations have certain time-dependent structures. In particular, they provide some recent developments on the empirical spectral distribution (ESD) of autocovariance matrices as well as the spectral analysis of low rank autoregressive coefficient matrices in a VAR model. These recent results can be seen as extensions of the ESD for sample covariance matrices and spectral analysis of spiked covariance matrices respectively in the i.i.d. setting. They also exhibit the additional inherent difficulties due to the time-dependence. The results can be further applied to solve many novel testing problems for autocovariance matrices and autoregressive coefficients in high-dimensional time series. We look forward to such fruitful results in their forthcoming papers.

For time-dependent observations, optimal estimation of structured covariance, precision and autocovariance matrices may also critically depend on the specific time series structures. Hence the techniques used in the analysis can be quite different from most problems considered in the present paper with i.i.d. observations. For example, in Section 2.2 of the present paper with $n = 1$, optimal estimation of autocovariance sequences in terms of the corresponding Toeplitz matrices is considered for one-dimensional stationary processes with p observations. The analysis, especially the lower bound argument, is very different from those for estimating other structured covariance and precision matrices. Recently, several extensions of the problems considered in the present paper to the high-dimensional time series were studied with various time-dependent structures. For example, besides those mentioned by Professors Paul and Wang, Chen et al. (2013) considered estimation of sparse covariance and precision matrices in a high-dimensional stationary process setting with autocovariance of any lag $k \geq 1$ treated as nuisance parameters. However, the fundamental difficulties of such matrix estimation problems in the time series settings remain largely unknown. New minimax lower bound techniques are much needed. It is of significant interest to consider optimal estimation with time-dependent observations.

5. Bandwidth selection for estimating bandable covariance matrices

Optimal selection of tuning parameters is of practical importance for a wide range of statistical problems including estimation of structured covariance matrices. Professor Zou raises an excellent question on optimal bandwidth selection for the tapering estimator $\hat{\Sigma}_{T,k}$ discussed in Section 2.1. This tuning parameter selection problem is especially relevant to estimating bandable covariance matrices. Indeed, as discussed in Section 2.1, theoretically optimal choices of the bandwidth k critically depend on the smoothness index α. They also vary significantly for various losses including the Frobenius norm, matrix ℓ_1 norm, and spectral norm losses. This is very different from estimating sparse covariance matrices in $\mathcal{H}(c_{n,p})$, in which optimal choices of the threshold in the proposed

thresholding estimator in Section 2.3 are not sensitive to the particular choice of loss functions considered in the paper or the sparsity parameter $c_{n,p}$. Professor Zou suggests using a SURE information criterion proposed in Yi and Zou (2013); Li and Zou (2014) to adaptively select the bandwidth by minimizing an unbiased risk estimator under the Frobenius norm loss. The selected bandwidth is shown to lead to a rate optimal estimator under the Frobenius norm loss without requiring the knowledge of α.

It is worthwhile to point out that the squared Frobenius norm is entrywise decomposable. As a consequence matrix estimation under the squared Frobenius norm loss essentially becomes a vector estimation problem and can be done in a row by row fashion. From this perspective, the success of the SURE approach is not very surprising. It is interest to see whether a similar procedure can be constructed for adaptive estimation under the spectral norm loss.

For the spectral norm loss, the adaptive procedure based on block thresholding discussed in the present paper is shown to be rate optimal. In particular, the choice of the thresholds is based on the spectral norm of each block matrix. This strategy can be adopted for estimation problems under other losses. For example, it can be shown that a Frobenius norm-based block thresholding estimator can adaptively achieve the optimal rate of convergence over the class $\mathcal{F}_\alpha(M_0, M)$ without knowing α.

6. Two-step sparse precision matrix estimation

For estimation of sparse precision matrices, Professor Zou points out a two-step approach proposed in Fan et al. (2014). This hybrid method combines the two major approaches discussed in the current paper, namely neighborhood-based CLIME and penalized likelihood approaches, and enjoys the merits of both methods. Specifically, this estimator is guaranteed to be positive semi-definite without an extra symmetrization step needed for CLIME. In addition, without the irrepresentibility condition typically required in penalized likelihood approaches, the hybrid method enjoys a strong oracle property as if the MLE with the knowledge of true support of the precision matrix $\Omega = (\omega_{ij})$. In particular, the estimator achieves the desired support recovery property. However, it seems that the strong oracle property relies on an extra assumption on the minimal magnitude of nonzero entries

$$\min_{i \neq j} |\omega_{ij}| \geq C \left(\sqrt{(\|\Omega\|_{\ell_1}^2 \log p)/n} \vee \sqrt{((1+K_3)^2 \log c_{n,p})/n} \right),$$

which can be seen in bounding the undesired probability in Theorem 3 of Professor Zou's discussion. Over the class $\mathcal{HP}(c_{n,p}, M)$ defined in (8), both the quantities $\|\Omega\|_{\ell_1}$ and K_3 are possibly diverging. Compared with the optimal ANT procedure stated in the Theorem 15 of the present paper, this hybrid estimator enjoys positive definiteness with the price of a stronger assumption.

This two-step philosophy is not unique: the CLIME estimator is calculated in the first step as the initializer to make the problem localizable such that

in the second step, the non-convex optimization is able to provide an optimal or oracle solution successfully. Indeed for many other non-convex approaches used in statistical problems, a relative good initial estimator is all needed to localize the problem. For example, for the sparse PCA problem, Birnbaum et al. (2013) established the two-step optimal procedure ASPCA with the help of the consistent estimator DT developed in Johnstone and Lu (2004) in their first step. For the sparse CCA problem, Gao et al. (2014) used a Fantope estimator in their first step of the two-step optimal procedure to successfully localize the CCA problem into a manageable regression model. We appreciate Professor Zou's insight and expect such philosophy can be applied to other problems in estimating structured covariance and precision matrices.

7. Miscellaneous

Professor Tsybakov makes the connection between the covariance matrix estimation problem and matrix estimation with additive noise $Y = \Sigma + W/\sqrt{n}$, where W is a random noise matrix with i.i.d. standard normal variables. These two problems are indeed closely related as the sufficient statistic for covariance matrix estimation, the sample covariance matrix $\hat{\Sigma}$ can be written as the same form $\Sigma + \hat{W}/\sqrt{n}$ although $\hat{W}$ has dependent sub-exponential entries. It would be very interesting to explore whether asymptotic equivalence between the two models in Le Cam's sense can be built rigorously.

Finally we thank Professor Tsybakov for the comment on Theorem 4 and for suggesting the alternative approach to sparse spiked covariance matrix estimation. Indeed, with a fixed bound on $\lambda_{n,p}$, the minimax rate under the squared spectral norm loss in Theorem 4 will be reduced to $\min\{1, c_{n,p}\log(ep/c_{n,p})/n\}$ since the rank $r_{n,p}$ cannot be larger than the sparsity $c_{n,p}$. In particular, the optimal rate does not depend on the rank. In such a setting, the alternative approach proposed by Professor Tsybakov has a simpler proof and provides a clearer interpretation. However, when allowing a growing $\lambda_{n,p}$, the rank $r_{n,p}$ plays a critical role in the minimax rate if $r_{n,p}\lambda_{n,p} \gg c_{n,p}\log(ep/c_{n,p})$. Specifically, the two terms in the optimal rate stated in Theorem 12 depend on $\lambda_{n,p}$ in different ways. To reveal this important phenomenon, it seems that the alternative approach cannot capture this subtlety and a more delicate procedure such as (15) is needed.

References

Barber, R. F. and M. Kolar (2015). ROCKET: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *arXiv preprint arXiv:1502.07641*.

Birnbaum, A., I. M. Johnstone, B. Nadler, and D. Paul (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *The Annals of Statistics 41*(3), 1055–1084. MR3113803

Cai, T. T. and A. Zhang (2015). Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. Technical report.

Chen, M., C. Gao, and Z. Ren (2015). Robust covariance matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691*.

Chen, X., M. Xu, and W. B. Wu (2013). Covariance and precision matrix estimation for high-dimensional time series. *The Annals of Statistics 41*(6), 2994–3021. MR3161455

Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics 42*(3), 819–849. MR3210988

Gao, C., Z. Ma, and H. H. Zhou (2014). Sparse CCA: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565*.

Han, F. and H. Liu (2013). Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *arXiv preprint arXiv:1305.6916*.

Johnstone, I. M. and A. Y. Lu (2004). Sparse principal components analysis. *Technical Report, Stanford University, Dept. of Statistics. arxiv.org, 0901.4392*.

Kolar, M. and E. Xing (2012). Consistent covariance selection from data with missing values. In J. Langford and J. Pineau (Eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, New York, NY, USA, pp. 551–558. Omnipress.

Li, D. and H. Zou (2014). Asymptotic properties of SURE information criteria for large covariance matrices. *arXiv preprint arXiv:1406.6514*.

Liu, H., F. Han, M. Yuan, J. Lafferty, and L. Wasserman (2012). High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics 40*(4), 2293–2326. MR3059084

Liu, H., F. Han, and C.-h. Zhang (2012). Transelliptical graphical models. In *NIPS*, pp. 809–817.

Loh, P.-L. and M. J. Wainwright (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics 40*(3), 1637–1664. MR3015038

Lounici, K. (2013). Sparse principal component analysis with missing observations. In *High Dimensional Probability VI*, pp. 327–356. Springer.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli 20*(3), 1029–1058. MR3217437

Mitra, R. and C.-H. Zhang (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. *arXiv preprint arXiv:1403.6195*.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, Volume 2, pp. 523–531. MR0426989

Wegkamp, M. and Y. Zhao (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli 22*(2), 1184–1226. MR3449812

Xue, L. and H. Zou (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics 40*(5), 2541–2571. MR3097612

Xue, L. and H. Zou (2013). Optimal estimation of sparse correlation matrices of semiparametric Gaussian copulas. *Statistics and Its Interface 7*, 201–209. MR3199378

Xue, L. and H. Zou (2014). Rank-based tapering estimation of bandable correlation matrices. *Statistica Sinica 24*, 83–100. MR3183675

Yi, F. and H. Zou (2013). SURE-tuned tapering estimation of large covariance matrices. *Computational Statistics & Data Analysis 58*, 339–351. MR2997947