COMMENTS on "MINIMAX ESTIMATION OF LARGE COVARIANCE MATRICES UNDER ℓ_1-NORM"

Ming Yuan

Georgia Institute of Technology

Professors Cai and Zhou are to be congratulated for making yet another important contribution to the development of theory and methodology for high-dimensional covariance matrix estimation. In this article, hereafter referred to as CZ, they considered large covariance matrix estimation under the matrix ℓ_1 loss for both sparse and bandable covariance matrices. As is common in the current literature, the results from CZ are derived under the subgaussian assumption as characterized by their (4). Thus far, it remains unknown how essential this assumption is. To partially address this intriguing question, I shall illustrate through a simple example that subgaussianity may not play a fundamental role in determining the difficulty of estimating a large covariance matrix.

Consider here the problem of estimating a large scale matrix for elliptically contoured distributions, a more general problem than estimating the covariance matrix for multiavariate normal distributions. Let $X \in \mathbb{R}^p$ have an elliptically contoured distribution in that there exist parameters $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ such that

$$X =_d \mu + rAU$$

where $r \geq 0$ is a random variable, U is uniformly distributed over the unit sphere in $\mathbb{R}^n$ and is independent of r, and $A \in \mathbb{R}^{p\times p}$ is a constant matrix such that $AA^\mathsf{T} = \Sigma$. In particular when r has a density, the density of X is

$$f(\mathbf{x}) = |\Sigma|^{-1/2} g((\mathbf{x}-\mu)^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\mu)), \qquad \mathbf{x} \in \mathbb{R}^p,$$

where g is the so-called kernel function uniquely determined by the distribution of r. Notable examples of elliptically contoured distribution are the multivariate normal, t, and the stable distributions. Note that many elliptically contoured distributions are not subgaussian and some do not even have finite second mo-

ments. For brevity, we assume that $\mu = 0$ and that Σ is a correlation-like matrix with ones on its diagonal. Our goal is to estimate Σ given a sample $X_1, \ldots, X_n$ consisting of independent copies of X. To fix ideas, wel focus on estimating sparse matrices. Write

$$\tilde{\mathcal{G}}_q(\rho, c_{n,p}) = \{\Sigma \in \mathcal{G}_q(\rho, c_{n,p}) : \Sigma_{ii} = 1 \quad \forall i\}.$$

Denote by $\mathcal{E}(\tilde{\mathcal{G}}_q(\rho, c_{n,p})$ the collection of centered elliptically contoured distributions with $\Sigma \in \tilde{\mathcal{G}}_q(\rho, c_{n,p})$. By the argument of CZ and Cai and Zhou (2011),

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{L}(X) \in \mathcal{E}(\tilde{\mathcal{G}}_q(\rho, c_{n,p})} \|\hat{\Sigma} - \Sigma\|^2 \gtrsim c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}, \tag{1.1}$$

where $\|\cdot\|$ is the matrix ℓ_α norm with any $\alpha \geq 1$. The question of interest here is whether or not this lower bound remains tight despite the lack of subgaussianity for many distributions from $\mathcal{E}(\tilde{\mathcal{G}}_q(\rho, c_{n,p})$. Interestingly, the answer is affirmative.

To this end, we need to construct a rate optimal estimator. We appeal to a useful property of elliptically contoured distributions. Let $Y = (Y_1, Y_2)^\mathsf{T}$ follow an elliptically contoured distribution with

$$\Sigma = \begin{pmatrix} 1 & \sigma \\ \sigma & 1 \end{pmatrix}.$$

Let $\tau = \mathbb{P}\{(Y_1 - Y_1^*)(Y_2 - Y_2^*) > 0\} - \mathbb{P}\{(Y_1 - Y_1^*)(Y_2 - Y_2^*) < 0\}$ be the population version of Kendall's τ statistic, where $Y^* = (Y_1^*, Y_2^*)^\mathsf{T}$ is an independent copy of Y. Then (see, e.g., Fang, Fang and Kotz (2002))

$$\tau = \frac{2}{\pi} \arcsin(\sigma).$$

Using this fact, we can estimate Σ in three steps.

(1) Estimate $\tau(X_i, X_j)$ by the sample Kendall's τ, denoted by $\hat{\tau}_{ij}$.

(2) Estimate Σ_{ij} by
$$\tilde{\Sigma}_{ij} = \sin\left(\frac{\pi}{2}\hat{\tau}_{ij}\right), \qquad \forall i \neq j.$$

(3) Let $\tilde{\Sigma}_{ii} = 1$ and apply thresholding to $(\tilde{\Sigma}_{ij})$:
$$\hat{\Sigma}_{ij} = \tilde{\Sigma}_{ij} I\left(\left|\tilde{\Sigma}_{ij}\right| \geq c\sqrt{\frac{\log p}{n}}\right)$$
for some numerical constant $c > 0$.

We argue that the resulting estimate $\hat{\Sigma}$ is indeed rate optimal. A careful examination of the proof of CZ reveals that it suffices to establish bounds for $|\tilde{\Sigma}_{ij} - \Sigma_{ij}|$ similar to their (24). This, as shown in Liu et al. (2012), can be achieved using Hoeffding's inequality for U-statistics. More specifically, we have

$$\mathbb{P}(|\tilde{\Sigma}_{ij} - \Sigma_{ij}| \geq t) \leq \exp\left(-\frac{nt^2}{2\pi^2}\right).$$

Using this in place of (24) of CZ, it can then be shown that

$$\sup_{\mathcal{L}(X)\in\mathcal{E}(\tilde{\mathcal{G}}_q(\rho,c_{n,p})} \|\hat{\Sigma} - \Sigma\|^2 \leq \sup_{\mathcal{L}(X)\in\mathcal{E}(\tilde{\mathcal{G}}_q(\rho,c_{n,p})} \|\hat{\Sigma} - \Sigma\|_1^2 \lesssim c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}. \quad (1.2)$$

Combining (1.1) and (1.2), we can conclude that

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{L}(X)\in\mathcal{E}(\tilde{\mathcal{G}}_q(\rho,c_{n,p})} \|\hat{\Sigma} - \Sigma\|^2 \asymp c_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-q}.$$

In this particular exercise, the subgaussian assumption is irrelevant. Of course, it is also a very specific example. The exact role of subgaussianity in high-dimensional covariance matrix estimation remains to be seen.

Acknowledgment

The research of Ming Yuan was supported in part by NSF Career Award DMS-0846234.

References

Cai, T.T. and Zhou, H. (2011) Optimal rates of convergence for sparse covariance matrix estimation, Technical Report.

Fang, H., Fang, K. and Kotz, S. (2002), The meta-elliptical distributions with fixed marginals, *Journal of Multivariate Analysis*, **82**, 1-16.

Liu, H., Han, F., Yuan, M. Lafferty, J. and Wasserman, L. (2012), High dimensional semiparametric Gaussian copula graphical models, Technical Report.

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

E-mail: myuan@isye.gatech.edu