Electronic Journal of Statistics
Vol. 10 (2016) 67–70
ISSN: 1935-7524
DOI: 10.1214/15-EJS1081A

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

Alexandre B. Tsybakov

Centre de Recherche en Economie et Statistique (UMR CNRS 9194), ENSAE
e-mail: alexandre.tsybakov@ensae.fr

Received January 2016.

Statistical inference for large covariance and precision matrices is a novel and interesting topic emerged in the last decade. The paper by Tony Cai, Zhao Ren and Harry Zhou (further referred to as [CRZ]) summarizes the key recent achievements in this rapidly developing area where the authors are among the leading contributors. The focus is on fundamental decision theoretic aspects, namely, on the following questions: (a) what are the best attainable rates of convergence of estimators in a minimax sense on various classes of matrices, and (b) how to construct data-driven adaptive procedures attaining these rates without the knowledge of the parameters of the classes. When the dimension of the covariance matrix is greater than the sample size, accurate estimation is problematic unless some assumptions are imposed on the structure of the matrix. A wealth of such structure assumptions is presented in the paper, most of them having the form of sparsity or approximate sparsity constraints. Sparsity here is understood either as a small number of non-zero entries or of non-zero columns/rows of the matrix, or as a small ℓ_q-norm of columns/rows, or as a low rank of the matrix, or as a combination of these properties.

The questions addressed in the paper have analogs in the classical Gaussian mean (Gaussian sequence) model, which is now extensively studied, cf., e.g., [2]. A key problem there is to construct minimax optimal and adaptive estimators of vectors on the ℓ_q-balls based on observation of the unknown vector in Gaussian noise. A straightforward matrix extension of this classical problem is estimation of a sparse matrix $\Sigma \in \mathbb{R}^{p\times p}$ from the observation

$$Y = \Sigma + \varepsilon W \tag{1}$$

where W is a random noise matrix with i.i.d. standard Gaussian entries and $\varepsilon > 0$ is the noise level that we can set as $\varepsilon = 1/\sqrt{n}$ in order to explore similarities

*Main article 10.1214/15-EJS1081.

with the covariance matrix estimation model. Some work about the minimax optimal estimation in model (1) under sparsity (in ordinary sense or in the sense of low rank) is now available, regarding mainly the estimation in the Frobenius norm (see, e.g., [3, 4, 8]). It would be interesting to see what are the differences or similarities with the covariance matrix estimation problem, under the same assumptions on Σ. Of course, for covariance matrix estimation, the model is somewhat different. Then, observations of the form (1) are also available with Y being the empirical covariance matrix but the noise matrix W is not Gaussian and its entries are not i.i.d. Furthermore, as compared to the above papers, there are more restrictions on matrix Σ since it s! hould be symmetric and positive definite. These differences make the analysis of covariance matrix estimation more involved, especially in what concerns the minimax lower bounds. However, intuitively it seems that there should be no fundamental difference in the rates between the two models. It would be interesting to clarify this point.

Consider one example, namely, the estimation of sparse spiked covariance matrices treated in Theorem 4 of [CRZ]. This theorem is based on a result in [1]. At first sight, it seems that the rate is different from what could be expected for model (1) under the same assumptions on Σ. Indeed, as shown in [4], the minimax rate of convergence under the spectral norm in model (1) does not depend on the rank, while the rank r appears in the rate of Theorem 4. However, Theorem 4, as well as its prototype in [1] are valid under the condition $r \le k$ where, $r = r_{n,p}$ and $k = c_{n,p}$ in the notation of [CRZ]. Thus, assuming that $\lambda_{n,p}$ is bounded by a fixed constant, we immediately deduce from Theorem 4 an upper bound of the order $k \log(ep/k)/n$ on the minimax risk. The lower bound is also of the same order. Therefore, with a fixed bound on $\lambda_{n,p}$, there is no dependency on the rank, which is in accordance with our initial guess based on the knowledge about model (1). The same result is easy to obtain by considering the estimator

$$\hat{\Sigma} = \operatorname*{argmin}_{\Sigma \in \mathcal{H}_0(k)} \|\Sigma^{\mathrm{emp}} - \Sigma\|_{(2k)}$$

where

$$\|A\|_{(2k)} = \max_{\|u\|_2=1, \|u\|_0 \le 2k} |u^T A u|,$$

Σ^{emp} denotes the empirical covariance matrix, and $\mathcal{H}_0(k)$ is the class of all covariance matrices of size $p \times p$ represented as $\Sigma = I + B$ with a symmetric matrix B having at most k non-zero rows and k non-zero columns. Here, $\|u\|_2$ is the Euclidean norm of $u \in \mathbb{R}^p$ and $\|u\|_0$ is the number of its non-zero components. Let $\Sigma^* \in \mathcal{H}_0(k)$ be the true covariance matrix. By definition of $\hat{\Sigma}$ we have

$$\|\Sigma^{\mathrm{emp}} - \hat{\Sigma}\|_{(2k)} \le \|\Sigma^{\mathrm{emp}} - \Sigma^*\|_{(2k)},$$

so that

$$\|\hat{\Sigma} - \Sigma^*\|_{(2k)} \le 2\|\Sigma^{\mathrm{emp}} - \Sigma^*\|_{(2k)}.$$

Since both $\hat{\Sigma}$ and Σ^* belong to $\mathcal{H}_0(k)$ the difference $\hat{\Sigma} - \Sigma^*$ has at most $2k$ non-zero rows and at most $2k$ non-zero columns. Thus, $\|\hat{\Sigma} - \Sigma^*\|_{(2k)} = \|\hat{\Sigma} - \Sigma^*\|$

where $\|\cdot\|$ denotes the spectral norm. On the other hand, if the observations X_i are i.i.d. $\mathcal{N}(0, \Sigma^*)$, and $2k \le n$, we have

$$\mathbb{E}\|\Sigma^{\text{emp}} - \Sigma^*\|^2_{(2k)} \le C\|\Sigma^*\|^2 \frac{k\log(ep/k)}{n} \le C(1+\lambda)^2 \frac{k\log(ep/k)}{n}, \qquad (2)$$

where λ is an upper bound on the spectral norm of B^* in the representation $\Sigma^* = I + B^*$, and $C > 0$ is an absolute constant. The first inequality in (2) follows from the results of [9, 5] and the union bound. In conclusion, we have

$$\sup_{\Sigma^* \in \mathcal{H}(k,\lambda)} \mathbb{E}\|\hat{\Sigma} - \Sigma^*\|^2 \le C(1+\lambda)^2 \frac{k\log(ep/k)}{n},$$

where $\mathcal{H}(k,\lambda) = \{\Sigma = I + B \in \mathcal{H}(k) : \|B\| \le \lambda\}$ is a larger class than the one considered in Theorem 4 of [CRZ], and this bound on the risk holds for any rank $r \le p$.

Another interesting point of comparison with model (1) arises in the context of missing data. When the dimensions are very high, assuming that all entries of the matrix are observed is often non-realistic. This motivated the theory of matrix completion, which is now a very elaborate field regarding mainly model (1). Much less is known about the behavior of estimators of covariance structures with missing data. First papers in this direction devoted to sparse PCA and to estimation of covariance matrices have appeared only very recently [6, 7]. The main question here is what is the largest fraction of missing values such that successful estimation of the matrix or of its caracteristics is still possible. The focus in [6, 7] is on the low rank covariance structures. The same question can be asked about various other covariance or precision matrix structures discussed in [CRZ].

References

[1] T. Cai, Z. Ma and Y. Wu (2015) Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, **161**, 781–815. MR3334281

[2] I.M. Johnstone (2015) *Gaussian estimation: Sequence and wavelet models.* Book draft.

[3] O. Klopp and A. B. Tsybakov (2015) Estimation of matrices with row sparsity. *Problems of Information Transmission*, **51**, 335–348.

[4] V. Koltchinskii, K. Lounici and A. B. Tsybakov (2011) Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, **39**, 1029–1058. MR2906869

[5] V. Koltchinskii and K. Lounici (2014) Concentration inequalities and moment bounds for sample covariance operators. To appear in *Bernoulli*. arxiv:1405.2468

[6] K. Lounici (2013) Sparse Principal Component Analysis with missing observations. In: *High dimensional probability VI*, volume 66 of Prog. Proba., *IMS Collections*, 327–356, Institute of Mathematical Statistics.

[7] K. Lounici (2014) High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, **20**, 2302–2329. MR3217437

[8] A. Rohde and A. B. Tsybakov (2011) Estimation of high-dimensional low rank matrices. *Annals of Statistics*, **39**, 887–930. MR2816342

[9] R. Vershynin (2012) Introduction to the non-asymptotic analysis of random matrices. In: *Compressed Sensing, Theory and Applications.* Edited by Y. Eldar and G. Kutyniok, Chapter 5, p. 210–268, Cambridge University Press. MR2963170