High dimensional probability 04/13/2022

Covariance estimation of sub-Gaussian distributions

Bowen Zhou *bowen_zz@sjtu.edu.cn*

1 Non-asymptotic results of sample covariance mareices

Suppose we observe n i.i.d. $p-$dimensional observations $\mathbf{X}_1, \cdots, \mathbf{X}_n$ distributed according to a sub-Gaussian distribution, with mean zero and covariance matrix $\mathbf{\Sigma}$. To estimate $\mathbf{\Sigma}$ we can use the sample covariance matrix:

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k \mathbf{X}_k^T.$$

Now we study the tail probability properties of this covariance estimation. The key point is the $\varepsilon - net$ argument.

Let $\mathcal{N}_\varepsilon$ is a ε-net of S^{p-1} with $|\mathcal{N}_\varepsilon| = \mathcal{N}(\mathrm{S}^{p-1}, \varepsilon)$. We know that

$$|\mathcal{N}_\varepsilon| \le \left(1 + \frac{2}{\varepsilon}\right)^p.$$

Now we deal with the operator norm of $\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$:

$$\begin{aligned}\mathbb{P}\left\{\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\| \ge t\right\} =& \mathbb{P}\left\{\sup_{x\in \mathrm{S}^{p-1}}\left|x^T\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right)x\right| \ge t\right\} \\ \le& \frac{1}{1-2\varepsilon}\mathbb{P}\left\{\sup_{x\in \mathcal{N}_\varepsilon}\left|x^T\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right)x\right| \ge t\right\} \\ \le& \frac{1}{1-2\varepsilon}\sum_{x\in \mathcal{N}_\varepsilon}\mathbb{P}\left\{\left|x^T\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right)x\right| \ge t\right\},\end{aligned}$$

then

$$x^T\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right)x = \frac{1}{n}\sum_{k=1}^{n} x^T\left(\mathbf{X}_k\mathbf{X}_k^T - \Sigma\right)x = \frac{1}{n}\sum_{k=1}^{n}\left(x^T\mathbf{X}_k\mathbf{X}_k^T x - x^T\Sigma x\right),$$

$x^T\mathbf{X}_k$ is sub-Gaussian due to $\mathbf{X}_k$ is sub-Gaussian, so $\left\{\left(x^T\mathbf{X}_k\mathbf{X}_k^T x - x^T\Sigma x\right), 1 \le k \le n\right\}$ are i.i.d. sub-exponential mean-zero random variables, so we can use Bernstein's inequality:

Lemma 1 (Bernstein's inequality)**.** *Let $X_1, \cdots, X_N$ be independent mean-zero sub-exponential random variables, and let $\mathbf{a} = (a_1, \cdots, a_N) \in \mathbb{R}^N$. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} a_i X_i\right| \ge t\right\} \le 2\exp\left(-c\min\left(\frac{t^2}{K^2\|a\|_2^2},\ \frac{t}{K\|a\|_\infty}\right)\right)$$

where $K = \max_i \|X_i\|_{\Psi_1}$.

Using this lemma, we have

$$\mathbb{P}\left\{\left|x^T\left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right)x\right| \ge t\right\} \le 2\exp\left(-c\min\left(\frac{t^2}{K^2},\ \frac{t}{K}\right)n\right),$$

where $K = \max_k \|x^T \mathbf{X}_k \mathbf{X}_k^T x - x^T \Sigma x\|_{\Psi_1} \leq C \max_k \|\mathbf{X}_k\|_{\Psi_2}^2 \leq C\|\Sigma\|$. Substituting it into the beginning, and let $\varepsilon = 1/4$, we have

$$\mathbb{P}\left\{\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\| \geq t\right\} \leq 2\exp\left(-c\min\left(\frac{t^2}{\|\Sigma\|^2},\ \frac{t}{\|\Sigma\|}\right)n + p\log 9\right), \tag{1.1}$$

let $t = \|\Sigma\| \max(\delta, \delta^2)$, then

$$\mathbb{P}\left\{\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\| \geq \|\Sigma\| \max(\delta, \delta^2)\right\} \leq 2\exp\big(-cn\delta^2 + p\log 9\big),$$

To counteract p and n in the exponential term, let $\delta = \tilde{C}\sqrt{p/n} + u\sqrt{p/n}$ and chose a large constant $\tilde{C}$ we have

$$\mathbb{P}\left\{\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\| \geq \|\Sigma\| \max(\delta, \delta^2)\right\} \leq 2\exp\big(-pu^2\big),$$

for all $u \geq 0$.

Remark. Suppose $p < n$, and recall the inequality (1.1), let $t = \tilde{C}\|\Sigma\|(\sqrt{p/n} + u)$, and choose a large enough constant $\tilde{C}$, we have

$$\mathbb{P}\left\{\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\| \geq \tilde{C}\|\Sigma\|\left(\sqrt{\frac{p}{n}} + u\right)\right\} \leq 2\exp\big(-c\min\big(u^2,\ u\big)n\big),$$

for all $u \geq 0$. From this we can see that

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\| = O_p\left(\sqrt{\frac{p}{n}}\right).$$

Remark. If we consider another norm of matrix: the max norm, i.e. $\|A\|_{\max} = \max_{ij}|A_{ij}|$, then what is the convergence order of covariance estimation? In fact,

$$\|A\|_{\max} = \max_{x,y\in\{e_k\}_{k=1}^n} \left|x^T A y\right|,$$

there are p^2 possible choice of (x, y) so with similar analysis of $\varepsilon - net$ argument, we have

$$\mathbb{P}\left\{\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\max} \geq t\right\} \leq 2\exp\left(-c\min\left(\frac{t^2}{\|\Sigma\|^2},\ \frac{t}{\|\Sigma\|}\right)n + 2\log p\right),$$

then by similar discussion, for all $u \geq 0$,

$$\mathbb{P}\left\{\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\max} \geq \|\Sigma\| \max(\delta, \delta^2)\right\} \leq 2\exp\big(-u^2\log p\big),$$

where $\delta = \tilde{C}\sqrt{\log p/n} + u\sqrt{\log p/n}$, and hence if $\log p < n$, then

$$\left\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_{\max} = O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

2 Sparse estimation in high dimension

In high dimension case, covariance matrices always have some additional structure. For example, some sparse structure may be a usual assumption.

Suppose the covariance matrix Σ is known to be sparse, but the positions of non-zero entries are no longer known.

2.1 Hard-thresholding estimator

Consider the $hard-thresholding$ estimator:

$$T_\lambda(\widehat{\mathbf{\Sigma}}) = \Big[\widehat{\mathbf{\Sigma}}_{ij}\mathbb{I}\Big(\widehat{\mathbf{\Sigma}}_{ij} \geq \lambda\Big)\Big].$$

Now we study the performance of this estimator, where the tuning parameter λ is suitably chosen, that is, we will show that

$$\Big\|T_\lambda(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\Big\| = O_p\left(\sqrt{\frac{\log p}{n}}\right).$$

Denote the adjacency matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ with entries $\mathbf{A}_{jk} = \mathbb{I}[\mathbf{\Sigma}_{jk} \neq 0]$. The following lemma are need to study the concentration inequality based on $\mathbf{A}$. In fact,

$$\Big\|T_\lambda(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\Big\| = O_p\left(\|A\|\sqrt{\frac{\log p}{n}}\right).$$

Lemma 2 (Non-negative matrices and operator norms)**.** *Given two $p-$dimensional symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, suppose that $0 \leq \mathbf{A}_{jk} \leq \mathbf{B}_{jk}$ for all $j, k = 1, \cdots, p$.*

(a) $0 \leq \mathbf{A}^m_{jk} \leq \mathbf{B}^m_{jk}$ *for all* $m = 1, 2, \cdots,\ j, k = 1, \cdots, p$.

(b) $\|A\| \leq \|B\|$.

(c) $\|C\| \leq \||C|\|$ *for any symmetric matrix C, where $|C|$ denotes the absolute value function applied elementwise.*

Proof of Lemma 2

$\square$

From section 1, we have, for all $u > 0$,

$$\mathbb{P}\left\{\Big\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\Big\|_{\max} \leq C\left(\sqrt{\frac{\log p}{n}} + u\right)\right\} \geq 1 - 2\exp\big\{-cn\min(u, u^2)\big\}.$$

Let $\lambda = C\left(\sqrt{\frac{\log p}{n}} + u\right)$. If $\Big\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\Big\|_{\max} \leq \lambda$, we are guaranteed that

$$\Big\|T_\lambda(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\Big\| \leq 2\|A\|\lambda.$$

First, $\Big\|T_\lambda(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\Big\| \leq \Big\|\Big|T_\lambda(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\Big|\Big\|$, and

$$\Big|T_\lambda(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\Big|_{ij} \leq \Big|T_\lambda(\widehat{\mathbf{\Sigma}}) - \widehat{\mathbf{\Sigma}}\Big|_{ij} + \Big|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\Big|_{ij} \leq 2\lambda A_{ij},$$

so by lemma 2, we have $\Big\|T_\lambda(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\Big\| \leq 2\|A\|\lambda$, and hence for all $u > 0$,

$$\mathbb{P}\left\{\Big\|T_\lambda(\widehat{\mathbf{\Sigma}}) - \mathbf{\Sigma}\Big\| \leq C\|A\|\left(\sqrt{\frac{\log p}{n}} + u\right)\right\} \geq 1 - 2\exp\big\{-cn\min(u, u^2)\big\}.$$

If the covariance matrix $\mathbf{\Sigma}$ has at most s non-zero entries per row, then

$$\|A\| \leq \sqrt{\|A\|_1\|A\|_\infty} \leq s,$$

and so

$$\left\| T_\lambda(\widehat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma} \right\| = O_p\left(s\sqrt{\frac{\log p}{n}} \right).$$

With similar argument, we can show that

$$\frac{1}{p}\left\| T_\lambda(\widehat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma} \right\|_F^2 = O_p\left(\frac{\|A\|_F^2}{p} \frac{\log p}{n} \right),$$

and if the covariance matrix $\boldsymbol{\Sigma}$ has at most s non-zero entries per row, then

$$\frac{1}{p}\left\| T_\lambda(\widehat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma} \right\|_F^2 = O_p\left(\frac{s \log p}{n} \right).$$

- **Approximate sparsity:** Given a parameter $q \in [0, 1]$ and a constant R_q, suppose $\boldsymbol{\Sigma}$ satisfies:

$$\max_i \sum_{j=1}^{p} |\boldsymbol{\Sigma}_{ij}|^q \leq R_q.$$

We will show that

$$\left\| T_\lambda(\widehat{\boldsymbol{\Sigma}}) - \boldsymbol{\Sigma} \right\| = O_p\left(R_q \left(\frac{\log p}{n} \right)^{(1-q)/2} \right).$$

2.2 Generalized thresholding estimator