Week 10
Spring 2009

Lecture 19. Estimation of Large Covariance Matrices: Upper bound

Observe

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \text{ i.i.d. from a } p\text{-variate Gaussian distribution, } N\left(\boldsymbol{\mu}, \Sigma_{p\times p}\right).$$

We assume that the covariance matrix $\Sigma_{p\times p} = (\sigma_{ij})_{1\le i,j\le p}$ is contained in the following parameter space,

$$\mathcal{F}(\alpha, \varepsilon, M) = \left\{ \begin{array}{c} \Sigma : |\sigma_{ij}| \le M\, |i-j|^{-(\alpha+1)} \text{ for all } k \\ \text{and } \lambda_{\max}(\Sigma) \le 1/\varepsilon \end{array} \right\} \tag{1}$$

Theorem 1 *Under the assumption (10), we have*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}} \mathbb{E} \left\| \hat{\Sigma} - \Sigma \right\|^2 \le C n^{-\frac{2\alpha}{2\alpha+1}} + C \frac{\log p}{n}. \tag{2}$$

This theorem tells us that there is an estimator $\hat{\Sigma}$ to obtain the rate $n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$. In the next lecture we will show this rate can not be improved. This result improves the rate $\left(\frac{\log p}{n}\right)^{\frac{\alpha}{\alpha+1}}$ in Theorem 1 in Bickel and Levina (2008a). When $\alpha = 1/2$ and $p = e^{\sqrt{n}}$, their rate is $n^{-\frac{1}{6}}$, while the rate in Theorem 5 is $n^{-\frac{1}{2}}$. The key reason for such an improvement is that we realized matrix estimation is fundamentally different from vector estimation!

Estimation Procedure:

Define

$$M_l^{(m)} = \left(\tilde{\sigma}_{ij} I\left\{l \le i < l+m, l \le j < l+m\right\}\right)_{p\times p}$$

and

$$S^{(m)} = \sum_{l=2-m}^{p} M_l^{(m)}$$

for all integers $2-m \le l \le p$ and $m \ge 1$. We estimate Σ by

$$\hat{\Sigma} = k^{-1}\left(S^{(2k)} - S^{(k)}\right). \tag{3}$$

We will set $k \asymp n^{\frac{1}{2\alpha+1}}$ for the operator norm and $n^{\frac{1}{2(\alpha+1)}}$ for the Frobenius norm.

Technically it is relatively easier to study this risk upper of the tapering estimator under the operator norm than the usual banding estimator.

Lemma 2 *We have*

$$\hat{\Sigma} = \left(w_{ij}\tilde{\sigma}_{ij}\right)_{p\times p}$$

where $w_{ij} = k^{-1}\{(2k - |i-j|)_+ - (k - |i-j|)_+\}$.

Note that

$$w_{ij} = k^{-1}\{(2k-|i-j|)_+ - (k-|i-j|)_+\} = \begin{cases} 1 & \text{when } |i-j| \le k \\ \in (0,1) & \text{when } k < |i-j| < 2k \\ 0 & \text{otherwise.} \end{cases}$$

Now we establish the risk upper bound for the estimator in equation (3) under the operator norm. We show that the variance part,

$$\mathbb{E}\left\|\hat{\Sigma} - \mathbb{E}\hat{\Sigma}\right\|^2 \le C\frac{k+\log p}{n} \tag{4}$$

and the bias part,

$$\left\|\mathbb{E}\hat{\Sigma} - \Sigma\right\|^2 \le Ck^{-2\alpha} \tag{5}$$

thus

$$\mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \le 2C\left(\frac{k+\log p}{n} + k^{-2\alpha}\right),$$

which implies

$$\mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|^2 \le 2C_1\left(n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}\right)$$

by setting

$$k = n^{\frac{1}{2\alpha+1}}. \tag{6}$$

Let C be a generic constant which may vary from place to place.

We prove the risk upper bound (5) for the bias part first. It is well known that the l_2 to l_2 norm of a symmetric matrix $A = (a_{ij})_{p\times p}$ is the bounded by its l_1 to l_1 norm, i.e.,

$$\|A\| \le \max_{i=1,\ldots,p} \sum_{j=1}^{p} |a_{ij}|.$$

We bound the operator norm of the bias part $\mathbb{E}\hat{\Sigma} - \Sigma$ by its l_1 to l_1 norm. Since $\mathbb{E}\tilde{\sigma}_{ij} = \sigma_{ij}$, we have

$$\mathbb{E}\hat{\Sigma} - \Sigma = ((w_{ij} - 1)\,\sigma_{ij})_{p\times p}$$

where $w_{ij} \in [0,1]$ and is exactly 1 when $|i-j| \le k$, then

$$\left\|\mathbb{E}\hat{\Sigma} - \Sigma\right\|^2 \le \left[\max_{i=1,\ldots,p} \sum_{j:|i-j|>k} |\sigma_{ij}|\right]^2 \le M^2 k^{-2\alpha}.$$

Now we establish (4) which is relatively complicated. The key idea in the proof is to write the whole matrix as an average of matrices which are sum of a large number of small disjoint block matrices, and for each small block matrix the classical random matrix theory can be applied. The following lemma shows that the operator norm of the random matrix $\breve{\Sigma} - \mathbb{E}\breve{\Sigma}$ is controlled by the maximum of operator norms of p number of $2k \times 2k$ random matrices. Let $M_l^{(m)} = (\tilde{\sigma}_{ij} I \{l \le i < l+m, l \le j < l+m\})_{p \times p}$. Define

$$N_l^{(m)} = \max_{1 \le l \le p-m+1} \left\| M_l^{(m)} - \mathbb{E} M_l^{(m)} \right\|.$$

Lemma 3 *Let $\breve{\Sigma}$ be defined as in (3). Then*

$$\left\| \breve{\Sigma} - \mathbb{E}\breve{\Sigma} \right\| \le 3 N_l^{(m)}.$$

For each small $m \times m$ random matrix with $m = 2k$, we control its operator norm as follows.

Lemma 4 *There is a constant $\rho_1 > 0$ such that*

$$\mathbb{P}\left\{ N_l^{(m)} > x \right\} \le 2p5^m \exp\left(-nx^2\rho_1\right) \tag{7}$$

for all $0 < x < \rho_1$ and $1 - m \le l \le p$.

With Lemmas 3 and 4 we are now ready to show the variance bound (4). By Lemma 3 we have

$$\begin{aligned} \mathbb{E}\left\| \breve{\Sigma} - \mathbb{E}\breve{\Sigma} \right\|^2 &\le 9\mathbb{E}\left(N_l^{(2k)}\right)^2 = 9\mathbb{E}\left(N_l^{(2k)}\right)^2 \left[I\left(N_l^{(2k)} \le x\right) + I\left(N_l^{(2k)} > x\right) \right] \\ &\le 9\left[x^2 + \mathbb{E}\left(N_l^{(2k)}\right)^2 I\left(N_l^{(2k)} > x\right) \right]. \end{aligned}$$

Note that $\left\| \mathbb{E}\breve{\Sigma} \right\| \le \|\Sigma\|$, which is bounded by a constant, and $\left\| \breve{\Sigma} \right\| \le \left\| \breve{\Sigma} \right\|_F$. The Cauchy–Schwarz inequality then implies

$$\begin{aligned} \mathbb{E}\left\| \breve{\Sigma} - \mathbb{E}\breve{\Sigma} \right\|^2 &\le C_1 \left[x^2 + \mathbb{E}\left(\left\| \breve{\Sigma} \right\|_F^2 + C \right) I\left(N_l^{(2k)} > x\right) \right] \\ &\le C_1 \left[x^2 + \sqrt{\mathbb{E}\left(\left\| \breve{\Sigma} \right\|_F + C \right)^4} \sqrt{\mathbb{P}\left(N_l^{(2k)} > x\right)} \right]. \end{aligned}$$

Set $x = 4\sqrt{\frac{\log p + m}{n\rho_1}}$. Then x is bounded by ρ_1 as $n \to \infty$. From Lemma 4 we obtain

$$\mathbb{E}\left\| \breve{\Sigma} - \mathbb{E}\breve{\Sigma} \right\|^2 \le C \left[\frac{\log p + m}{n} + p^2 \cdot \left(p5^m \cdot p^{-8} e^{-8m}\right)^{1/2} \right] \le C_1 \left(\frac{\log p + m}{n} \right). \quad \blacksquare \tag{8}$$

Now we give proofs of auxiliary lemmas.

Proof of Lemma 2: It is easy to see

$$\begin{aligned} k w_{ij} &= \#\{l : \{i,j\} \subset \{l, \ldots, l+2k-1\}\} - \#\{l : (i,j) \subset \{l, \ldots, l+k-1\}\} \\ &= (2k - |i-j|)_+ - (k - |i-j|)_+, \end{aligned}$$

which takes value in $[0,k]$. Clearly from the above, $kw_{ij}=k$ for $|i-j|\le k$.

Proof of Lemma 3: Without loss of generality we assume that p can be divided by m. Set $\delta_l^{(m)}=M_l^{(m)}-\mathbb{E}M_l^{(m)}$. By (3)

$$\left\|S_l^{(m)}-\mathbb{E}S_l^{(m)}\right\|\le\sum_{l=1}^{m}\left\|\sum_{-1\le j\ <\ p/m}\delta_{jm+l}^{(m)}\right\|. \tag{9}$$

Since $\delta_{jm+l}^{(m)}$ are diagonal blocks of their sum over $-1\le j<p/m$, we have

$$\left\|S_l^{(m)}-\mathbb{E}S_l^{(m)}\right\|\le m\max_{1\le l\le m}\left\|\sum_{0\le j\ <\ p/m}\delta_{jm+l}^{(m)}\right\|\le m\max_{2-m\le l\le p}\left\|\delta_l^{(m)}\right\|.$$

This and (3) imply the conclusion, since $\delta_l^{(k)}$ and $\delta_l^{(2k)}$ are all sub-blocks of certain matrix $\delta_l^{(2k)}$ with $1\le l\le p-2k+1$.

Proof of Lemma 4 : For any $m\times m$ symmetric matrix A, we have

$$\left|u^TAu\right|-\left|v^TAv\right|\le\left|u^TAu-v^TAv\right|=\left|(u-v)^T A\,(u+v)\right|\le\|u-v\|\,\|A\|\,\|u+v\|$$

Let $S_{1/2}^{m-1}$ be a 1/2 net of the unit sphere S^{m-1} in the Euclidean distance in R^m. We have

$$\|A\|\le\sup_{u\in S^{m-1}}\left|u^TAu\right|\le\sup_{u\in S_{1/2}^{m-1}}\left|u^TAu\right|+\frac{1}{2}\|A\|\frac{3}{2}=\sup_{u\in S_{1/2}^{m-1}}\left|u^TAu\right|+\frac{3}{4}\|A\|$$

which implies $\|A\|\le 4\sup_{u\in S_{1/2}^{m-1}}\left|u^TAu\right|$. Since we are allowed to pack $\mathrm{Card}\left(S_{1/2}^{m-1}\right)$ balls of radius 1/4 into a $1+1/4$ ball in R^m, volume comparison yields

$$(1/4)^m\mathrm{Card}\left(S_{1/2}^{m-1}\right)\le(5/4)^m,$$

i.e., $\mathrm{Card}\left(S_{1/2}^{m-1}\right)\le 5^m$. Thus there exist $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_{5^m}\in S^{m-1}$ such that

$$\|A\|\le 4\sup_{j\le 5^m}\left|v_j^TAv_j\right|,\ \text{for all } m\times m \text{ symmetric } A.$$

This one step approximation argument is similar to the proof of Proposition 4.2 (ii) in Zhang and Huang (2008).

Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ be i.i.d. p-vectors with $\mathbb{E}\left(\mathbf{X}_1-\boldsymbol{\mu}\right)\left(\mathbf{X}_1-\boldsymbol{\mu}\right)^T=\Sigma$. Under the Gaussian (sub-Gaussian) assumption there exists $\rho>0$ such that

$$\mathbb{P}\left\{\mathbf{v}^T(\mathbf{X}_i-\mathbb{E}\mathbf{X}_i)(\mathbf{X}_i-\mathbb{E}\mathbf{X}_i)^T\mathbf{v}>x\right\}\le e^{-x\rho/2}\ \text{for all } x>0 \text{ and } \|\mathbf{v}\|=1$$

which implies $\mathbb{E}\left(t\mathbf{v}^T(\mathbf{X}_i-\mathbb{E}\mathbf{X}_i)(\mathbf{X}_i-\mathbb{E}\mathbf{X}_i)^T\mathbf{v}\right)<\infty$ for all $t<\rho/2$ and $\|\mathbf{v}\|=1$, then there exists $\rho_1>0$ such that

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}^T\left[(\mathbf{X}_i-\mathbb{E}\mathbf{X}_i)(\mathbf{X}_i-\mathbb{E}\mathbf{X}_i)^T-\Sigma\right]\mathbf{v}\right|>x\right\}\le e^{-nx^2\rho_1/2}$$

for all $0 < x < \rho_1$ and $\|\mathbf{v}\| = 1$. (See, e.g., Chapter 2 in Saulis and Statulevicius (1991).) Thus we have

$$\begin{aligned}
\mathbb{P}\left\{\max_{1\le l\le p-m+1}\left\|M_l^{(m)} - \mathbb{E}M_l^{(m)}\right\| > x\right\} &\le \sum_{1\le l\le p-m+1} \mathbb{P}\left\{\left\|M_l^{(m)} - \mathbb{E}M_l^{(m)}\right\| > x\right\} \\
&\le 2p5^m \sup_{\mathbf{v}_j, l} \mathbb{P}\{|\mathbf{v}_j^T(M_l^{(m)} - \mathbb{E}M_l^{(m)})\mathbf{v}_j| > x\} \\
&\le 2p5^m \exp\left(-nx^2\rho_1/2\right). \quad \blacksquare
\end{aligned}$$

Remark: The proof here works for sub-Gaussian assumption which is slightly more general than Gaussian.

Lecture 19. Estimation of Large Covariance Matrices: Lower bound (I)

Observe

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \text{ i.i.d. from a } p\text{-variate Gaussian distribution, } N\left(\boldsymbol{\mu}, \Sigma_{p\times p}\right).$$

We assume that the covariance matrix $\Sigma_{p\times p} = (\sigma_{ij})_{1\le i,j\le p}$ is contained in the following parameter space,

$$\mathcal{F}(\alpha,\varepsilon,M) = \left\{\Sigma : |\sigma_{ij}| \le M\,|i-j|^{-(\alpha+1)} \text{ for } i\neq j \text{ and } \lambda_{\max}(\Sigma) \le 1/\varepsilon\right\}. \tag{10}$$

In addition we assume that $p \le e^n$. In this lecture we will see this assumption is necessary to estimate $\Sigma_{p\times p}$ consistently under the operator norm.

Theorem 5 *Under the assumption (10), we have*

$$\inf_{\hat\Sigma}\sup_{\mathcal{F}} \mathbb{E}\left\|\hat\Sigma - \Sigma\right\|^2 \ge cn^{-\frac{2\alpha}{2\alpha+1}} + c\frac{\log p}{n}. \tag{11}$$

In this lecture we will show

$$\inf_{\hat\Sigma}\sup_{\mathcal{F}} \mathbb{E}\left\|\hat\Sigma - \Sigma\right\|^2 \ge c\frac{\log p}{n}$$

by using Le Cam's method. Next time we will apply Assouad's lemma to prove the other part of the lower bound

$$\inf_{\hat\Sigma}\sup_{\mathcal{F}} \mathbb{E}\left\|\hat\Sigma - \Sigma\right\|^2 \ge cn^{-\frac{2\alpha}{2\alpha+1}}.$$

We will apply Le Cam's method to derive a lower bound for minimax risk. Let X be an observation from a distribution in the collection $\{P_\theta, \theta\in\Theta = \{\theta_0,\theta_1,\ldots,\theta_p\}\}$. Le Cam's method, which is based on a two-point testing argument, gives a lower bound for the maximum estimation risk over the parameter set Θ. More specifically, let L be the loss function. Define $r(\theta_0,\theta_m) = \inf_t\left[L(t,\theta_0) + L(t,\theta_m)\right]$ and $r_{\min} = \inf_{1\le m\le p} r(\theta_0,\theta_m)$, and denote $\bar{\mathbb{P}} = \frac{1}{p}\sum_{m=1}^{p}\mathbb{P}_{\theta_m}$.

Lemma 6 *Let T be an estimator of θ based on an observation from a distribution in the collection $\{P_\theta, \theta\in\Theta = \{\theta_0,\theta_1,\ldots,\theta_p\}\}$, then*

$$\sup_\theta \mathbb{E}L(T,\theta) \ge \frac{1}{2} r_{\min}\left\|\mathbb{P}_{\theta_0}\wedge\bar{\mathbb{P}}\right\|$$

For $1\le m\le p_1$, let Σ_m be a diagonal covariance matrix with $\sigma_{mm} = 1 + \sqrt{\tau\frac{\log p_1}{n}}$, $\sigma_{ii} = 1$ for $i\neq m$, and let Σ_0 be the identity matrix. Let $\mathbf{X}_l = \left(X_1^l, X_2^l, \ldots, X_p^l\right)^T \sim N(0,\Sigma_m)$, and denote the joint density of $\mathbf{X}_1,\ldots,\mathbf{X}_n$ by f_m, $1\le m\le p_1$ with $p_1 = \max\{p, \exp(n/2)\}$, which can be written as follows

$$f_m = \prod_{1\le i\le n, 1\le j\le p, j\neq m}\phi_1\left(x_j^i\right)\cdot\prod_{1\le i\le n}\phi_{\sigma_{mm}}\left(x_m^i\right)$$

where ϕ_σ, $\sigma = 1$ or σ_{mm}, is the density of $N\left(0, \sigma^2\right)$. Let $\theta_m = \Sigma_m$ for $0 \leq m \leq p_1$ and the loss function L be the squared operator norm. It is easy to see $d\left(\theta_0, \theta_m\right) = \frac{1}{2}\tau \frac{\log p_1}{n}$ for all $1 \leq m \leq p_1$. Then the lower bound (**??**) follows immediately if there is a constant $c > 0$ such that

$$\left\|\mathbb{P}_{\theta_0} \wedge \bar{\mathbb{P}}\right\| \geq c. \tag{12}$$

Since $\int q_0 \wedge q_1 d\mu = 1 - \frac{1}{2}\int |q_0 - q_1|\, d\mu$ for any two densities q_0 and q_1, and the Jensen's inequality implies

$$\left[\int |q_0 - q_1|\, d\mu\right]^2 = \left(\int \left|\frac{q_0 - q_1}{q_1}\right| q_1 d\mu\right)^2 \leq \int \frac{(q_0 - q_1)^2}{q_1} d\mu = \int \frac{q_0^2}{q_1} d\mu - 1.$$

Hence $\int q_0 \wedge q_1 d\mu \geq 1 - \frac{1}{2}\left(\int \frac{q_0^2}{q_1} d\mu - 1\right)^{1/2}$. To establish equation (12), it thus suffices to show that $\int \left(\frac{1}{p_1}\sum_{m=1}^{p_1} f_m\right)^2 / f_0 d\mu - 1 \to 0$, i.e.,

$$\int \frac{1}{p_1^2}\sum_{m=1}^{p_1} \frac{f_m^2}{f_0} d\mu + \frac{1}{p_1^2}\sum_{m \neq j} \frac{f_m f_j}{f_0} d\mu - 1 \to 0. \tag{13}$$

We now calculate $\int \frac{f_m f_j}{f_0} d\mu$. For $m \neq j$ it is easy to see

$$\int \frac{f_m f_j}{f_0} d\mu - 1 = 0.$$

When $m = j$, we have

$$\begin{aligned} \int \frac{f_m^2}{f_0} d\mu &= \frac{\left(\sqrt{2\pi\sigma_{mm}}\right)^{-2n}}{\left(\sqrt{2\pi}\right)^{-n}} \prod_{1 \leq i \leq n} \int \exp\left[\left(x_m^i\right)^2 \left(-\frac{1}{\sigma_{mm}} + \frac{1}{2}\right)\right] dx_m^i \\ &= \left[1 - (1 - \sigma_{mm})^2\right]^{-n/2} = \left(1 - \tau\frac{\log p_1}{n}\right)^{-n/2}. \end{aligned}$$

Thus

$$\begin{aligned} \int \left(\frac{1}{p_1}\sum_{m=1}^{p_1} f_m\right)^2 / f_0 d\mu - 1 &= \frac{1}{p_1^2}\sum_{m=1}^{p_1}\left(\int \frac{f_m^2}{f_0} d\mu - 1\right) \leq \frac{1}{p_1}\left(1 - \tau\frac{\log p_1}{n}\right)^{-n/2} - \frac{1}{p_1} \\ &= \exp\left[-\log p_1 - \frac{n}{2}\log\left(1 - \tau\frac{\log p_1}{n}\right)\right] - \frac{1}{p_1} \to 0 \end{aligned} \tag{14}$$

for $0 < \tau < 1$, where the last step follows from the inequality $\log(1 - x) \geq -2x$ for $0 < x < 1/2$.