S&DS 411 (Fall 2019) Selected Topics in Statistical Decision Theory **Lecture: 10**

Structured Covariance and Precision Matrices Estimation

Lecturer: Harrison Zhou *Scribe: Cynthia Yue, Nov 6, 2019*

1 Introduction

Let us consider the model

$$X_1, \ldots, X_n \text{ i.i.d.} \quad \mathcal{N}(0, \sum_{p\times p}) \quad \text{where each } X_i \text{ is a vector.}$$

The goal is to estimate the covariance $\sum_{p\times p}$. We may first think that we've already been estimating vectors, and the covariance is just a big vector, so we might as well just treat the covariance matrix as a vector. However, because it is a matrix, we can estimate $\sum_{p\times p}$ differently. Specifically, our goal is to estimate $\sum_{p\times p}$ under different losses:

$$\underbrace{\|\widehat{\sum} - \sum\|_F}_{\text{Frobenius norm}}, \quad \underbrace{\|\widehat{\sum} - \sum\|_2}_{\text{spectral norm}}, \quad \|\widehat{\Omega} - \Omega\|.$$

Let us first cover some review:

1. Spectral norm $\|A\|_2 = \sup_{\|x\|_2 \le 1} \|Ax\|_2$. (largest of all the eigenvalues)
   Note: If we have good control of the spectral norm, we will also have good control of the eigenvalues and eigenvectors. Hence, if we do inference and we're also interested in eigenvalues, the spectral norm is good.
2. We denote the precision matrix Ω as $\sum^{-1}$, the inverse of the covariance. High precision means a big Ω and a small $\sum$, which is why we call it the precision matrix. For our data, we prefer to have small $\sum$. The Gaussian graphical model estimates the precision matrix.
3. $\mathbb{E}X_1X_1^T = \sum$.
4. With Davis-Kahan, we have the following: If we are considering the spectral norm, the question we have in mind is what we will say about the first eigenvalue. In other words, what is the upper band? Let's say we have

   $$\widehat{\sum} \text{ eigenvalues } \widehat{\lambda}_1 \ge \ldots \ge \widehat{\lambda}_p$$
   $$\sum \text{ eigenvalues } \lambda_1 \ge \ldots \ge \lambda_p. \quad \text{(population)}$$

   Then, we have

   $$|\widehat{\lambda}_1 - \lambda_1| \le \|\widehat{\sum} - \sum\|_2.$$

   This is why the spectral norm is good.

5. If we have $\sum$: first principal component $\underbrace{v_1}_{\substack{\text{the eigenvector} \\ \text{associated with the} \\ \text{largest eigenvalue}}}$ with eigenvalues $\lambda_1 > \lambda_2$, the spectral norm is small, and $\widehat{\sum}$: first principal component $\widehat{v}_1$, then what can we say about $\|v_i - \widehat{v}_i\|$?

$$\|v_1 - \widehat{v}_1\|_2 \lesssim \frac{\|\widehat{\sum} - \sum\|_2}{\lambda_1 - \lambda_2}.$$

This means that we have good control over the principal components.

2 Structured Covariance Matrices Estimation

The idea with structured covariance matrices estimation is that we do not have a lot of data (n), but we have lots of columns (p), so we need to put in structure. Many applications can be found in financial econometrics.

2.1 Bandable Covariance

$$\sum = \begin{pmatrix} & & & & & & \overbrace{\quad}^{\text{nearly 0}} \\ \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} & \dots & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \dots & \sigma_{3k+1} & \dots & \sigma_{2p} \\ \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \dots & \dots & \dots & \sigma_{pp} \end{pmatrix}.$$

$\mathbb{E}_{x_{1i}x_{1j}}$ small when $|i-j|$ big

Bandable means that if we have

$$X_1 = \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1p} \end{pmatrix},$$

the covariance between X_{11} and X_{13} will be smaller than the covariance between X_{11} and X_{12}. If two entries are close, the covariance is large; if they are far, the covariance is small.

From Bicket and Levina (2004), we have the following assumptions:

$$|\sigma_{ij} \leq c|i-j|^{-\alpha-1}, \quad i \neq j \text{ and } \alpha > 0;$$
$$|\sigma_{ii}| \leq M.$$

The idea is that when i and j are far from each other—when the distance is larger—then the dependence is smaller. In other words, when σ is further from the diagonal, then will be smaller, but along the diagonal, will be larger.

For simplicity, we also assume $|\sigma_{ii}| \leq M$.

$$\begin{aligned}
\widetilde{\sum} &= \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \\
\widetilde{\sigma_{12}} &= \frac{1}{n}\sum_{i=1}^{n} X_{i1}X_{i2} \\
&\approx \mathcal{N}\left(\sigma_{12}, \frac{1}{n}(\sigma_{11}\sigma_{22}+\sigma_{12}^2)\right) \\
\mathbb{E}|\widetilde{\sigma_{ij}} - \sigma_{ij}| &\le \frac{1}{\sqrt{n}}.
\end{aligned}$$

Our estimation procedure is to start with MLE and method of moments.

$$\widehat{\sigma}_{ij} = \begin{cases} \widetilde{\sigma_{ij}} & |i-j| \le k \\ 0 & |i-j| > k. \end{cases}$$

The hard part is how to control the spectral norm.

$$(\widehat{\sum} - \sum)_{ij} = \begin{cases} \widetilde{\sigma}_{ij} - \sigma_{ij} & |i-j| \le k \\ \underbrace{-\sigma_{ij}}_{\text{this is the bias}} & |i-j| > k. \end{cases}$$

Let's find the spectral norm of this.

$$\begin{aligned}
\|\widehat{\sum} - \sum\|_2 &\le \|A\|_2 + \|B\|_2 \quad \text{from triangle inequality} \\
\widehat{\sum} - \sum &= A + B \\
A = (a_{ij}), \quad a_{ij} &= \begin{cases} \widetilde{\sigma}_{ij} - \sigma_{ij} & |i-j| \le k \\ 0 & \text{otherwise} \end{cases} \\
\underbrace{B}_{\text{bias}} = (b_{ij}), \quad b_{ij} &= \begin{cases} 0 & |i-j| \le k \\ -\sigma_{ij} & \text{otherwise.} \end{cases}
\end{aligned}$$

Specifically, we have

$$\|\widehat{\sum} - \sum\|_2 \lesssim \frac{k}{\sqrt{n}}\sqrt{\log p} + k^{-\alpha} \lesssim \left(\frac{\log p}{n}\right)^{\frac{\alpha}{2(1+\alpha)}} \quad \text{where } k^{1+\alpha} = \sqrt{\frac{n}{\log p}}.$$

How do we find the spectral norm of B?

We have the fact

$$\|A\|_2 \le \sup_i \left(\sum_j |a_{ij}|\right).$$

Using this fact, we get

$$\begin{aligned}\|B\|_2 &\leq \sup_i \sum_{j:|j-i|\geq k} |\sigma_{ij}| \\ &\leq 2c(k^{-\alpha-1} + (k+1)^{-\alpha-1} + \ldots) \\ &\leq \underbrace{c'k^{-\alpha}}_{\text{upper bound for the bias}}.\end{aligned}$$

If $\alpha = 0$, this upper bound goes to ∞, which is why we have the assumption $\alpha > 0$.

Now, what about A?

$$\|A\|_2 \leq \sup_i \sum_{j:|j-i|\leq k} |\widetilde{\sigma}_{ij} - \sigma_{ij}| \leq c' \frac{k}{\sqrt{n}} \sqrt{\log p} \quad \text{with high probability.}$$

A good question now is if this is rate optimal.

Now, using the Bickel approach to bound the spectral norm with the L1 norm, what upper bound do we get?

Remark. 1. With Bickel and Levina, we have

$$\|\widetilde{\sum} - \sum\|_2 \leq \frac{p}{\sqrt{n}} \sqrt{\log p}.$$

2. In classical, random matrix theory, what is the upper bound?

$$\|\widetilde{\sum} - \sum\|_2 \leq c\left(\sqrt{\frac{p}{n}} + \frac{Z}{\sqrt{n}}\right) \quad \text{where } Z \sim \mathcal{N}(0,1).$$

This is a deep result in random matrix theory.

From this, we see that the Bickel and Levina approach is not really optimal.

We need a better analysis for optimality, which will be inspired by classical theory with random matrices. If we look at the $k \times k$ submatrix, we have $\|A\| \lesssim \sqrt{\frac{k}{n}}$ with high probability, but we may have multiple $k \times k$ submatrices, so we have

$$\|A\|_2 \lesssim \sqrt{\frac{k}{n}} + \sqrt{\frac{\log p}{n}}.$$

Claim.

$$\|\widehat{\sum} - \sum\| \lesssim \sqrt{\frac{k}{n}} + k^{-\alpha} + \sqrt{\frac{\log p}{n}} \lesssim n^{-\frac{\alpha}{2\alpha+1}} + \sqrt{\frac{\log p}{n}} \quad \text{where } k = n^{\frac{1}{1+2\alpha}}.$$

This is rate optimal, and Fano's Lemma shows optimality.

2.2 Sparse Covariance

$$\sum = (\sigma_{ij})_{1\leq i,j\leq p}.$$

We have the following assumptions:

$$|\sigma_{ii}| \leq M \quad \text{for all } i.$$

For every column, look at the nonzero entries and that are $\leq s$:

$$\sup_j \underbrace{Card\{i \neq j; \sigma_{ij} \neq 0\}}_{\|\sigma_{\cdot j}\|_0} \leq s.$$

Remark.

$$\sigma_{ij} = 0 \Leftrightarrow \mathbb{E}X_{1i}X_{1j} = 0.$$

In the Gaussian case, this means that they are independent.

If we look at our sample covariance, we have

$$\widetilde{\sum} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$$
$$\widetilde{\sigma}_{ij} \approx \mathcal{N}(\sigma_{ij}, \frac{1}{n}(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2)).$$

Remark.

$$|\frac{1}{\sqrt{n}}(\widetilde{\sigma}_{ij} - \sigma_{ij})| \leq \frac{c}{\sqrt{n}}(1 + Z^2)$$
$$\text{where } Z \sim \mathcal{N}(0,1) \text{ and } |Z| \leq \epsilon\sqrt{n}.$$

So what is our procedure? There are a lot of zeros. What are we going to do? Our procedure for estimation is that it is like sparse vector estimation, so we will do something like LASSO or simpler:

$$\widehat{\sigma}_{ij} = \begin{cases} \widetilde{\sigma}_{ij} & |\widetilde{\sigma}_{ij}| \geq \lambda \leftarrow \text{thresholding procedure} \\ 0 & \text{otherwise.} \end{cases}$$

Remark.

$$\max_{i,j} \frac{|\widetilde{\sigma}_{ij} - \sigma_{ij}|}{\sqrt{\sigma ii \sigma_{jj} + \sigma_{ij}^2}} \leq \sqrt{\frac{2\log p^2}{n}} = 2\sqrt{\frac{\log p}{n}} \quad \text{with high probability.}$$

This is similar to LASSO thresholding.

What is the risk upper bound?

Claim.

$$\|\widehat{\sum} - \sum\|_2 \leq \sup_i \sum_j |\widehat{\sigma}_{ij} - \sigma_{ij}|$$
$$\lesssim \sqrt{\frac{\log p}{n}} s \quad \text{with high probability.}$$

This is slightly bigger than $\frac{1}{\sqrt{n}}$ because we picked a big threshold. The rate is optimal when $s \lesssim p^\beta$ and $\beta < \frac{1}{2}$, or, in other words, when s is small.

3 Precision Matrices Estimation

3.1 Bandable Precision Matrix

$$\Omega = (\omega_{ij}).$$

From autoregression in time series, we have assumptions:

$$|\omega_{ij}| \leq c|i-j|^{-\alpha-1} \quad \text{for } i \neq j$$
$$0 < m \leq |\omega_{ii}| \leq M < +\infty.$$

Claim.

$$\|\hat{\Omega} - \Omega\|_2 \lesssim n^{-\frac{\alpha}{2\alpha+1}} + \sqrt{\frac{\log p}{n}} \quad \text{with high probability.}$$

This is different from the covariance matrix because, with covariance, we have sample covariance. With the precision matrix, when $p > n$, how do we invert the covariance? This is very difficult!

3.2 Sparse Precision Matrices Estimation

We have the assumptions:

$$\sup_j \|\omega_{\cdot j}\|_0 \leq s$$
$$0 < M \leq \omega_{ii} \leq M < +\infty.$$

We use the fact

$$\omega_{12} = 0 \Leftrightarrow X_{11} \perp X_{12} | X_{13}, \ldots, X_{1p} \quad \text{conditional independence.}$$

4 Matrix Functionals Estimation

Let us consider the model:

$$X_1, \ldots, X_n \text{ i.i.d.} \quad \mathcal{N}(0, \sum).$$

We have the assumptions:

$$\sum = \sum_{k=1}^{r} \lambda_k v_k v_k^T$$
$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r.$$

Let us consider the special case: $r = 1$. Our goal is to estimate v_1, but if we put sparsity $\|v_1\|_0 \leq s$ and $\lambda_1 \geq c$, how can we estimate $\sum$?

Claim.

$$\|\widehat{\sum} - \sum\|_2 \lesssim s\sqrt{\frac{\log p}{n}}$$
$$\|\hat{v}_1 - v_1\| \lesssim s\sqrt{\frac{\log p}{n}} \quad \text{from Davis-Kahan.}$$

Is this rate optimal? Can we get $\sqrt{\frac{s \log p}{n}}$ rate? Because we have s parameters, ideally we want $\lesssim \sqrt{\frac{s \log p}{n}}$). Can we do it? It turns out that it depends on s.