Week 9
Spring 2009

Lecture 17. Assouad's Lemma and minimax lower bound for functional linear regression

The Assouad's lemma gives a lower bound for the maximum risk over the parameter set $\Lambda = \{0,1\}^r$, in an abstract form, applicable to the problem of estimating an arbitrary quantity $\psi(\gamma)$, belonging to a metric space space with metric d. Let $H(\gamma,\gamma') = \sum_{i=1}^r |\gamma_i - \gamma_i'|$ be the Hamming distance on $\{0,1\}^r$, which counts the number of positions at which γ and γ' differ. In this lecture we will apply this lemma to the functional linear regression.

Assouad's Lemma. For any estimator T based on an observation in the model $\{\mathbb{P}_\gamma, \gamma \in \Lambda\}$, and any $p > 0$

$$\max_\gamma 2^p \mathbb{E}_\gamma d^p(T, \psi(\gamma)) \geq \min_{H(\gamma,\gamma')\geq 1} \frac{d^p(\psi(\gamma), \psi(\gamma'))}{H(\gamma,\gamma')} \frac{r}{2} \min_{H(\gamma,\gamma')=1} \|\mathbb{P}_\gamma \wedge \mathbb{P}_{\gamma'}\|.$$

Functional linear regression
Assume that data pairs $(Y_i, X_i(t))$ for $i = 1, 2, \cdots, n$ are i.i.d. with

$$Y_i = a + \int_0^1 b(t) X_i(t) dt + \xi_i \quad 1 \leq i \leq n \tag{1}$$

where $X_i(t)$'s are i.i.d. Gaussian processes and $\xi_i \sim N(0,1)$. The main task is to estimate the slope function $b(t)$.

The distribution of a gaussian process $X(t)$ is uniquely determined by its mean process $\mu(t) = \mathbb{E}X(t)$ and covariance kernel $K(s,t) = \mathbb{E}Z(s)Z(t)$, where $Z(t) = X(t) - \mu(t)$. If the covariance kernel K is in $L^2([0,1]^2)$, it has a L^2-spectral decomposition,

$$K(s,t) = \sum_{j=1}^{+\infty} \theta_j \phi_j(s) \phi_j(t) \tag{2}$$

By convention, the eigenvalues are arranged in decreasing order, $\theta_1 \geq \theta_2 \geq \cdots \geq 0$. The eigenfunctions $\phi_1, \phi_2, \cdots$ form a complete orthonormal basis of $L^2([0,1])$ of real-valued functions that are square integrable with respect to Lebesgue measure on $[0,1]$. Note that the contribution from $\mu(t)$ can be absorbed into the intercept, so that (1) becomes

$$Y_i = b_0 + \int_{\mathcal{T}} b(t) Z_i(t) dt + \xi_i, \quad \text{with} \quad b_0 = a + \int_{\mathcal{T}} b(t)\mu(t) dt. \tag{3}$$

Condition 1 *Let $\beta > 0$ and $M_i > 0$ for $i = 0, 1$. Define the function class for b by*

$$b = \sum_{j=1}^{\infty} b_j \, \phi_j, \text{ with } |b_j| \leq M_1 \, j^{-\beta}, \text{ for all } j = 1, 2, ... \tag{4}$$

We can interpret this as a "smoothness class" of functions, where the functions become "smoother" (measured in the sense of generalized Fourier expansions in the basis $\{\phi_j\}$) as β increases. We shall also assume the eigenvalues satisfy

$$M_0^{-1} \, j^{-\alpha} \leq \theta_j \leq M_0 \, j^{-\alpha} \tag{5}$$

Let $\mathcal{F}(\alpha, \beta, M_0, M_1)$ denote the set of distributions F of (X, Y) that satisfies (4) and (5).

Theorem 2 *Under the condition above we have*

$$\inf_{\hat{b}} \sup_{F \in \mathcal{F}(\alpha,\beta,M_0,M_1)} \mathbb{E} \int_{\mathcal{T}} (\hat{b}(t) - b(t))^2 dt \geq c n^{-(2\beta-1)/(\alpha+2\beta)}$$

for some c depending on α, β, M_0 and M_1.

Proof. We first define a subset $\mathcal{F}_n$ of $\mathcal{F}(\alpha, \beta, M_0, M_1)$. Let $a = 0$, $\mu(t) \equiv 0$, and the covariance kernel $K_0(s,t) = \sum_{j \geq 1} \theta_j \phi_j(s) \phi_j(t)$ and the eigenvalues $\theta_j = j^{-\alpha}$, for $j \geq 1$. Let

$$b_\gamma(t) = \sum_{L_n < j \leq 2L_n} c_1 j^{-\beta} \gamma_j \phi_j(t).$$

where $L_n = c_0 n^{1/(\alpha+2\beta)}$. Note that there is a one to one correspondence between $\mathcal{F}_n$ and $\mathcal{W}_n = [0,1]^{L_n}$. It is easy to verify $\mathcal{F}_n \subset \mathcal{F}$.

By Assouad's lemma with $p = 2$, it follows that

$$\max_{\gamma \in \mathcal{W}_n} \mathbb{E} \int_{\mathcal{T}} (\hat{b}(t) - b(t))^2 dt \geq \frac{c_1^2}{2} \min_{h(\gamma,\gamma') \geq 1} \frac{\sum_{j \in W_n} j^{-2\beta} (\gamma_j - \gamma'_j)^2}{h(\gamma, \gamma')} \cdot L_n \cdot \min_{h(\gamma,\gamma')=1} ||\mathbb{P}_\gamma \wedge \mathbb{P}_{\gamma'}|| \tag{6}$$

It is easy to see

$$\min_{h(\gamma,\gamma') \geq 1} \frac{\sum_{j \in W_n} j^{-2\beta} (\gamma_j - \gamma'_j)^2}{h(\gamma, \gamma')} = \min_{h(\gamma,\gamma') \geq 1} \frac{\sum_{j \in W_n} j^{-2\beta} (\gamma_j - \gamma'_j)^2}{\sum_{j \in W_n} (\gamma_j - \gamma'_j)^2} \geq (2L_n)^{-2\beta} \tag{7}$$

If we can show that

$$\min_{h(\gamma,\gamma')=1} ||\mathbb{P}_\gamma \wedge \mathbb{P}_{\gamma'}|| \geq c_2 \tag{8}$$

then

$$\max_{\gamma \in \mathcal{W}_n} \mathbb{E} \int_{\mathcal{T}} (\hat{b}(t) - b(t))^2 dt \geq c {L_n}^{-(2\beta-1)} = c n^{-(2\beta-1)/(\alpha+2\beta)}.$$

We know

$$||\mathbb{P}_\gamma \wedge \mathbb{P}_{\gamma'}|| \geq \frac{1}{2}\alpha_2^2\left(\mathbb{P}_\gamma, \mathbb{P}_{\gamma'}\right) = \left(1 - \frac{1}{2}\mathbf{H}^2(\mathbb{Q}_\gamma, \mathbb{Q}_{\gamma'})\right)^{2n}$$

where $\mathbb{Q}_\gamma$ is the joint distribution of one single copy of (Y, X) with parameter γ. Note that

$$\begin{aligned}\mathbf{H}^2(\mathbb{Q}_\gamma, \mathbb{Q}_{\gamma'}) &\leq c_3 \int_0^1 \int_0^1 \left[b_\gamma(s) - b_{\gamma'}(s)\right]\left[b_\gamma(t) - b_{\gamma'}(t)\right] K_0(s,t) ds dt \\ &= c_4 j^{-\alpha-2\beta} = c_4/n,\end{aligned}$$

then equation (8) follows immediately.

Remark 3 *This approach can be applied to many functional regression models such as generalized functional linear regression and single index model.*

Lecture 18. Estimation of Large Covariance Matrices: Introduction

Observe that

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

For simplicity, we assume $\boldsymbol{\mu}$ is 0. The maximum likelihood estimator is

$$\tilde{\Sigma} = \frac{1}{n}\sum_{l=1}^{n} \mathbf{X}_l \mathbf{X}_l^T$$

for $n \geq p$ and write $\tilde{\Sigma} = (\tilde{\sigma}_{ij})_{1\leq i,j\leq p}$. Let $\mathbf{X}_l = \left(X_1^l, X_2^l, \ldots, X_p^l\right)^T$. We then write

$$\tilde{\sigma}_{ij} = \frac{1}{n}\sum_{l=1}^{n} X_i^l X_j^l.$$

Let $\Sigma_{p\times p} = (\sigma_{ij})_{1\leq i,j\leq p}$. It is then easy to see

$$\mathbb{E}\tilde{\sigma}_{ij} = \sigma_{ij} \tag{9}$$

$$\mathbb{V}ar(\tilde{\sigma}_{ij}) = \frac{1}{n}\left(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2\right) \tag{10}$$

i.e., $\tilde{\sigma}_{ij}$ is an unbiased estimator of σ_{ij} with a variance $\left(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2\right)/n$.

Following Bickel and Levina (2008a) we assume that the covariance matrix $\Sigma_{p\times p} = (\sigma_{ij})_{1\leq i,j\leq p}$ is contained in the following parameter space,

$$\mathcal{F}(\alpha, \varepsilon, M) = \left\{\Sigma : |\sigma_{ij}| \leq M|i-j|^{-(\alpha+1)} \text{ for all } i \neq j \text{ and } \lambda_{\max}(\Sigma) \leq 1/\varepsilon\right\}. \tag{11}$$

In addition, let's assume that $p \geq \gamma n$ for some $\gamma > 0$. If we see a matrix $A = (a_{ij})_{p\times p}$ as a vector with with length p^2, the Frobenius norm of a matrix $A = (a_{ij})_{p\times p}$ is just the l_2 norm of this vector and so defined as follows

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}.$$

It is easy to see that the operator norm is bounded by the Frobenius norm, i.e., $\|A\| \leq \|A\|_F$. The following theorem gives the minimax rate of convergence under the Frobenius norm.

Theorem 4 *Under the assumption (11), we have*

$$\inf_{\hat{\Sigma}} \sup_{\mathcal{F}} \mathbb{E}\frac{1}{p}\left\|\hat{\Sigma} - \Sigma\right\|_F^2 \leq Cn^{-\frac{2\alpha+1}{2(\alpha+1)}}. \tag{12}$$

Proof. Since the entries of the covariance matrix decay to 0 as moving away from the diagonal, Bickel and Levina (2008a) naturally propose to use a banding estimator

$$\hat{\Sigma}_{banding} = \left(\tilde{\sigma}_{ij} I\left\{|i-j| \leq k\right\}\right)_{p\times p}.$$

We will estimate each row or column separately under the square l_2 loss by choosing an appropriate k. The method was applied for years in nonparametric estimation in orthogonal bases regression. Since

$$\begin{aligned} \mathbb{E}\tilde{\sigma}_{ij} &= \sigma_{ij} \\ \mathbb{V}ar(\tilde{\sigma}_{ij}) &= \mathbb{E}\left(\xi_{ij}\right)^2 = \frac{1}{n}\left(\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2\right), \end{aligned}$$

we have

$$\frac{1}{p}\mathbb{E}\left\|\hat{\Sigma} - \Sigma\right\|_F^2 \leq \frac{1}{p}\sum_{\{(i,j):k<|i-j|\}} \sigma_{ij}^2 + \frac{1}{p}\sum_{\{(i,j):|i-j|\leq k\}} \frac{\sigma_{ii}\sigma_{jj} + \sigma_{ij}^2}{n} = R_1 + R_2$$

The assumption $\lambda_{\max}(\Sigma) \leq 1/\varepsilon$ implies that $\sigma_{ii} \leq 1/\varepsilon$ for all i. Since $|\sigma_{ij}|$ is uniformly bounded for all i and j, we immediately have $R_2 \leq C\frac{k}{n}$. Now we show that

$$\frac{1}{p}\sum_{\{(i,j):k<|i-j|\}} \sigma_{ij}^2 \leq Ck^{-2\alpha-1}$$

which is apparently true for $|\sigma_{ij}| \leq C_1 |i-j|^{-(\alpha+1)}$ for all $i \neq j$.

$$\mathbb{E}\frac{1}{p}\left\|\hat{\Sigma} - \Sigma\right\|_F^2 \leq Ck^{-2\alpha-1} + C\frac{k}{n} \leq C_2 n^{-\frac{2\alpha+1}{2(\alpha+1)}} \tag{13}$$

by choosing k appropriately,

$$k = n^{\frac{1}{2(\alpha+1)}}. \tag{14}$$

■

Remark: The choice of k here is different from the optimal choice for the operator norm which will be discussed next time.

For a matrix A let's define

$$\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2.$$

This is often called this operator norm. More precisely it is an l_2 to l_2 norm of matrix A. When A is symmetric, it is known that $\|A\|$ is equal to the magnitude of the largest eigenvalue of A. Hence it is also called spectral norm. It is well known that the operator norm of a symmetric matrix $A = (a_{ij})_{p\times p}$ is bounded by its l_1 norm, i.e.,

$$\|A\| \leq \|A\|_1 = \max_{i=1,\ldots,p} \sum_{j=1}^{p} |a_{ij}|.$$

This fact can be argued easily as follows. Let λ be an eigenvalue of A, and $v = (v_i)_{1\le i\le p}$ be a corresponding eigenvector, i.e., $Av = \lambda v$. Let $|v_i| = \|v\|_\infty$, and write $\lambda = \sum_{j=1}^p a_{ij}\frac{v_j}{v_i}$, then we have $|\lambda| \le \sum_{j=1}^p |a_{ij}|\left|\frac{v_j}{v_i}\right| \le \sum_{j=1}^p |a_{ij}| \le \max_{i=1,\ldots,p}\sum_{j=1}^p |a_{ij}|$. Bickel and Levina (2008) showed the following result.

Theorem 5 *Under the assumption (11), we have*

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

Denote that $\hat\Sigma - \mathbb{E}\hat\Sigma$ by $V = (v_{ij})$. Note that Bickel and Levina (2008) controlled the operator norm by the l_1 to l_1 norm as follows

$$\begin{aligned}\mathbb{E}\left\|\hat\Sigma - \mathbb{E}\hat\Sigma\right\|^2 &\le \mathbb{E}\left\|\hat\Sigma - \mathbb{E}\hat\Sigma\right\|_1 = \mathbb{E}\left(\max_{j=1,\ldots,p}\sum_i |v_{ij}|\right)^2 \\ &\le C\left(\frac{k}{\sqrt{n}}\sqrt{\log p}\right)^2 = C\frac{k^2\log p}{n}\end{aligned}$$

Note that $\mathbb{E}\sum_i |v_{ij}| \le Ck/\sqrt{n}$. It is then expected that $\mathbb{E}\left(\max_{j=1,\ldots,p}\sum_i |v_{ij}|\right)^2 \le C\left(\frac{k}{\sqrt{n}}\sqrt{\log p}\right)^2$ (see their paper for details) and so

$$\mathbb{E}\left\|\hat\Sigma - \Sigma\right\|_1^2 \le C\frac{k^2\log p}{n} + Ck^{-2\alpha}$$

An optimal tradeoff of k is then $\left(\frac{\log p}{n}\right)^{\frac{1}{2(\alpha+1)}}$ which implies a rate of $\left(\frac{\log p}{n}\right)^{-\frac{\alpha}{\alpha+1}}$ in Theorem 1 in Bickel and Levina (2008). It is much slower than the rate $n^{-\frac{2\alpha}{2\alpha+1}} + \frac{\log p}{n}$ we will obtain later.