ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 18: Density estimation and Structured Estimation

Lecturer: Yihong Wu *Scribe: Yuheng Bu, Apr. 7, 2016*

18.1 Estimator based on pairwise comparison: LeCam-Birgé

The idea of constructing estimator based on pairwise tests is due to Le Cam ([LC86], see also [vdV02, Section 10]) and Birgé [Bir83]. We are given n i.i.d. samples $X_1, X_2, \cdots, X_n$ generated from P, where $P \in \mathcal{P}$ is the density to be estimated. Let the loss function between the true distribution P and the estimated distribution $\hat{P}$ be their Hellinger distance, i.e.

$$\ell(P, \hat{P}) = H^2(P, \hat{P}).$$

Then, we have the following results.

Theorem 18.1 (Le Cam-Birgé)**.**

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P H^2(P, \hat{P}) \lesssim \epsilon_n^2,$$

where

$$n\epsilon_n^2 \asymp \log N_H(\mathcal{P}, \epsilon_n^2)$$

and $N_H(\mathcal{P}, \epsilon_n^2)$ is the covering number of set $\mathcal{P}$ under Hellinger distance.

Note: Recall from Fano's inequality, we have a minimax lower bound ϵ_n^2 with KL divergence loss which satisfies

$$\log N_{KL}(\mathcal{P}, \epsilon_n^2) \asymp n\epsilon_n^2.$$

Proof. Let $P_1, \cdots, P_N$ be the maximal ϵ-packing of $\mathcal{P}$ under Hellinger distance. Thus, $\forall i \neq j$,

$$H(P_i, P_j) \geq \epsilon,$$

and for $\forall P \in \mathcal{P}$, $\exists i \in [N]$, s.t.

$$H(P, P_i) \leq \epsilon,$$

Consider the following **Pairwise test problem**, we choose $i \neq j$, and testing between two ϵ balls:

$$\begin{cases} H_0 : P \in B(P_i, \epsilon) \\ H_1 : P \in B(P_j, \epsilon) \end{cases} \quad i \neq j, \text{ s.t. } \quad H(P_i, P_j) \geq \delta = 3\epsilon.$$

Thus, we know that $\forall P \in B(P_i, \epsilon)$, $\forall Q \in B(P_j, \epsilon)$,

$$H(P, Q) \geq H(P_i, P_j) - 2\epsilon = \delta - 2\epsilon = \epsilon.$$

Suppose we have an optimal test ψ_{ij}, and $\psi_{ij} = 0$ corresponding to H_0, $\psi_{ij} = 1$ corresponding to H_1. For this optimal test, we have the following large deviation bound,

$$\sup_{P \in B(P_i, \epsilon)} \mathbb{P}(\psi_{ij} = 1) \leq \exp(-\frac{n}{2}\epsilon^2),$$

$$\sup_{P \in B(P_j, \epsilon)} \mathbb{P}(\psi_{ij} = 0) \leq \exp(-\frac{n}{2}\epsilon^2).$$

We construct our density estimator as follows. For $i \in [N]$, consider,

$$T_i = \begin{cases} \max_{i\in[N]} H^2(P_i, P_j) & \text{s.t. } \psi_{ij} = 1.H(P_i, P_j) > \delta; \\ 0, & \text{no such } j \text{ exists.} \end{cases}$$

Basically, T_i records the maximum distance that the optimal test ϕ_{ij} will confuse between P_i and P_j from the true underlying distribution P, with the constraint $H(P_i, P_j) > \delta$. And our estimator is set to be

$$\hat{P} = P_{i^*}, \quad \text{where} \quad i^* \in \arg\min_{i\in[N]} T_i.$$

In our analysis, we assume that $P \in B(P_1, \epsilon)$. Typically,

$$T_1 = 0,\ T_j \geq \delta^2,\ \forall\ j \text{ s.t.} \quad H(P_1, P_j) \geq \delta.$$

So the probability that the loss function is bigger than 4ϵ can be bounded as,

$$\begin{aligned} \mathbb{P}(H(\hat{P}, P) > 4\epsilon) &\leq \mathbb{P}(i^* \in \{j : H(P_1, P_j) > \delta\}) \\ &\leq \mathbb{P}(T_1 > 0), \end{aligned}$$

where

$$\begin{aligned} \mathbb{P}(T_1 > 0) &= \mathbb{P}(\exists\ j,\ H(P_1, P_j) > \delta \text{ and } \psi_{1j} = 1) \\ &\leq N(\epsilon) \sup_{j:H(P_1,P_j)>\delta} \mathbb{P}(\ \psi_{1j} = 1) \\ &\leq N(\epsilon) \exp(-\frac{n\epsilon^2}{2}). \end{aligned}$$

We used the union bound and the large deviation bound in the last inequality. Here, we can see if we choose $n\epsilon^2 \asymp \log N_H(\mathcal{P}, \epsilon^2)$, the probability that the bias is bigger than ϵ is bounded by a exponential bound, thus we can conclude that $\sup_{P\in\mathcal{P}} \mathbb{E}_P H^2(P, \hat{P}) \lesssim \epsilon_n^2$. □

Remark 18.1. The result on $N(\epsilon)$ can be improved by using local metric entropy (doubling), which means using the different radius balls for different j to pack the set $\mathcal{P}$.

18.2 Structured estimation problem

Let's begin this section with our favourite example, the Gaussian Location model.

Example 18.1 (GLM)**.** Consider the p-dimensional n-sample GLM. We have

$$Y_i = \theta + Z_i,$$

where $\theta \in \Theta \subseteq \mathbb{R}^p$, $i \in [n]$. We have n i.i.d copies of Y, and the noise $Z \sim \mathcal{N}(0, I_p)$. We consider the quadratic minimax loss for this estimation problem,

$$R_n^*(\Theta) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2.$$

Here, we add some structure for the parameter space Θ to study the so called **denoising by sparsity** problem, let

$$\Theta = \{\text{all } k\text{-sparse vectors}\} = B_0(k) = \{\theta \in \mathbb{R}^p, \|\theta\|_0 \le k\}, \quad k \in [p],$$

where $\|\theta\|_0 = |\{i : \theta_i \neq 0\}|$ is the number of nonzero entries of θ, indicating the sparsity of θ. We want to analysis the asymptotic behavior of $R_n^*(B_0(k))$.

Remark 18.2. The set $B_0(k)$ can be written as a union of linear subspace of $\mathbb{R}^p$.

$$B_0(k) = \bigcup_{S\subseteq[p],|S|\le k} \{\theta, \theta_{S^C} = 0\}.$$

Remark 18.3. To study the behavior of $R_n^*(B_0(k))$, it is sufficient to consider one sample and the risk $R_1^*(B_0(k))$. Indeed, we have

$$R_n^*(B_0(k)) = \frac{1}{n} R_1^*(B_0(k)).$$

Proof. Since $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ is the sufficient statistics of this problem, and $\bar{Y} \sim \mathcal{N}(\theta, \frac{1}{n} I_p)$. Given n i.i.d. samples, it is sufficient to solve the following one-dimensional problem,

$$\bar{Y} = \theta + \frac{1}{\sqrt{n}} Z \quad \Leftrightarrow \quad \sqrt{n}\bar{Y} = \sqrt{n}\theta + Z,$$

where $Z \sim \mathcal{N}(0, I_p)$. Since $\sqrt{n}B_0(k) = B_0(k)$, estimating $\sqrt{n}\theta$ has the same minimax risk for estimating θ given one sample. Thus,

$$\begin{aligned} R_1^*(B_0(k)) &= \inf_{\hat{\theta}} \sup_{\sqrt{n}\theta\in\sqrt{n}B_0(k)} \mathbb{E}_\theta \|\sqrt{n}\hat{\theta} - \sqrt{n}\theta\|_2^2 \\ &= n \inf_{\hat{\theta}} \sup_{\theta\in B_0(k)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \\ &= n R_n^*(B_0(k)). \end{aligned}$$

Thus, in the following discussion, we only consider the case of $n = 1$. □

Claim 18.1 (The Oracle Lower bound)**.** *Note that for $n = 1$, given the information of the position of nonzero entries, we have the following lower bound which has been proved before*

$$R_1^*(B_0(k)) \ge k.$$

Theorem 18.2. *Actually, the minimax risk for this sparsity problem is*

$$R_1^*(B_0(k)) \asymp k + \log \binom{p}{k} \asymp k \log \frac{ep}{k}.$$

Note: If $k = o(p)$ and $p \to \infty$, we have the following results (proved in homework)

$$R_1^*(B_0(k)) = \big(2 + o(1)\big) k \log \frac{p}{k}.$$

18.2.1 Lower bound for denoising by sparsity

We will prove the lower bound in 18.2 by mutual information method. Consider the following binary sequences set:

$$B = \{b \in \{0,1\}^p : w_H(b) = k\},$$

where $w_H(b)$ is the Hamming weights of b. Suppose that b is drawn uniformly from the set B, and $\theta = \tau b$. Here $\tau > 0$ and will be specified later. Thus, we have the following Markov chain which represents our problem model,

$$b \to \theta \to Y \to \hat{\theta} \to \hat{b}.$$

Denote the set $G = \tau B$, so $\theta \in G$. The mutual information is upper bounded by the radius of set G,

$$\begin{aligned} I(\theta;\hat{\theta}) \le I(\theta;Y) &\le \mathrm{rad}_{kL}(\mathcal{N}(\theta, I_p), \theta \in G) \\ &\le \sup_{\theta \in G} D(P_\theta \| P_0) \\ &= \sup_{\theta \in G} \frac{1}{2}\|\theta\|_2^2 = \frac{k\tau^2}{2}. \end{aligned}$$

To give a lower bound for $I(\theta;\hat{\theta})$, we need a bound for $\min_{\mathbb{E}\|\theta-\hat{\theta}\|_2^2} I(\theta;\hat{\theta})$. Consider

$$\hat{b} = \arg\min_{b \in B} \|\hat{\theta} - \tau b\|_2^2.$$

Since $\hat{b}$ is the minimizer of $\|\hat{\theta} - \tau b\|_2^2$, we have,

$$\|\tau\hat{b} - \theta\|_2 \le \|\tau\hat{b} - \hat{\theta}\|_2 + \|\theta - \hat{\theta}\|_2 \le 2\|\theta - \hat{\theta}\|_2.$$

Thus,

$$\tau^2 d_H(b,\hat{b}) = \|\tau\hat{b} - \theta\|_2^2 \le 4\|\theta - \hat{\theta}\|_2^2,$$

where d_H denotes the Hamming distance between b and $\hat{b}$. Suppose that $\mathbb{E}\|\hat{\theta} - \theta\| \le \epsilon\tau^2 k$, then we have $\mathbb{E}d_H(b,\hat{b}) \le 4\epsilon k$, and

$$\begin{aligned} I(\hat{b};b) &\ge \min_{\mathbb{E}d_H(b,\hat{b}) \le 4\epsilon k} I(\hat{b};b) \\ &= H(b) - \max_{\mathbb{E}d_H(b,\hat{b}) \le 4\epsilon k} H(b|\hat{b}) \\ &= \log \binom{p}{k} - \max_{\mathbb{E}d_H(b,\hat{b}) \le 4\epsilon k} H(b - \hat{b}|\hat{b}) \\ &\ge \log \binom{p}{k} - \max_{\mathbb{E}d_H(b,\hat{b}) \le 4\epsilon k} H(b - \hat{b}). \end{aligned}$$

Since this optimization problem has the solution[1],

$$\max_{\mathbb{E}w_H(W) = m, W \in \{0,1\}^p} H(W) = ph(\frac{m}{p}),$$

[1]It can be easily verified that the maximum is achieved with the distribution $\mathrm{Bern}(\frac{m}{p})^{\otimes p}$, write this distribution as $q(w) = (\frac{m}{p})^{w_H(w)}(1-\frac{m}{p})^{p-w_H(w)}$. For any $p(w)$ satisfies $\mathbb{E}(w_H(W)) = m$, we have $H(W) = -D(p\|q) + \mathbb{E}_p[\log\frac{1}{q(w)}] \le \mathbb{E}_p[\log\frac{1}{q(w)}] = m\log\frac{p}{m} + (p-m)\log\frac{p}{p-m} = ph(\frac{m}{p})$.

where $h(\alpha) = -\alpha \log \alpha - \bar{\alpha} \log \bar{\alpha}$ is the binary entropy function. If $\alpha < \frac{1}{4}$, $h(\alpha) \asymp -\alpha \log \alpha$. For $\epsilon = \frac{1}{16}$ and let $k \leq \frac{p}{10}$, then $4\epsilon k/p \leq 1/4$, we can use the asymptotic results for $h(\alpha)$. Combine this with the previous bound, we get

$$I(\hat{b}; b) \geq \log \binom{p}{k} - ph(\frac{4\epsilon k}{p}) \asymp k \log \frac{p}{k}.$$

Thus, we have the following bound by mutual information method,

$$k \log \frac{p}{k} \lesssim I(\theta; Y) \lesssim k\tau^2.$$

Remember we choose $R^* = \epsilon\tau^2 k$, thus we can conclude that

$$R^* = \epsilon\tau^2 k \gtrsim k \log \frac{p}{k}.$$

Combining with the result in the oracle lower bound, we have

$$R^* \gtrsim k + k \log \frac{p}{k}.$$

Note: This problem can not be solved by letting each coordinate of θ to be i.i.d. Bernoulli random variable, i.e. $\theta_i \in \text{Bern}(\frac{k}{P})$. Since in this case, $\|\theta\|_0 \sim \text{Binomial}(p, \frac{k}{p})$. As for large p and k, we know that the Binomial distribution can be approximated by a Poisson distribution. With a constant probability $\text{Poi}(1) > 1$, thus θ is not k sparse.

Remark 18.4. For the case $k = o(p)$, we can show

$$R^*_{k,p} \geq (2 + o_p(1))k \log \frac{p}{k}.$$

To prove this result, we need to first show that for the case $k = 1$,

$$R^*_{1,p} \geq (2 + o_p(1)) \log p.$$

Next, show that for any k, the minimax risk is lower bounded by the Bayesian risk with the block prior. The block prior is that we divide the p-coordinate into k blocks, and pick one coordinate from each p/k-coordinate uniformly. With this prior, one can show

$$R^*_{k,p} \geq kR^*_{1,p/k} \asymp k \log \frac{p}{k}.$$

18.2.2 Upper bound for denoising by sparsity

In this subsection, we will prove the upper bound for 18.2. We will use the following results on the maxima of Gaussian, proved in our homework.

$$Y = \theta + Z, \quad Z \sim \mathcal{N}(0, I_p),$$

then,

$$\|Z\|_\infty \leq \sqrt{2 \log p} + o_p(1).$$

Given this result, it is natural to consider the following minimization problem,

- ℓ_0-minimization
$$\hat{\theta} = \arg\min \|\theta\|_0, \quad \text{s.t.} \quad \|y - \theta\|_\infty \le \tau = \sqrt{2\log p}.$$
- ℓ_1-minimization
$$\hat{\theta} = \arg\min \|\theta\|_1, \quad \text{s.t.} \quad \|y - \theta\|_\infty \le \tau = \sqrt{2\log p}.$$

However, the estimator given by these two constraint minimization problem only can show

$$\sup_{\|\theta\|_0 \le k} \mathbb{E}_\theta \|\theta - \hat{\theta}\| \lesssim k \log p,$$

which does not meet the desired result. Thus, we will look at the Maximum Likelihood estimator. For Gaussian distribution,

$$P_\theta(y) \propto \exp\left(-\frac{\|y-\theta\|_2^2}{2}\right).$$

Thus, the MLE is equivalent to the minimum distance rule,

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\|\theta\|_0 \le k} \|y - \theta\|_2^2.$$

We can show for this constraint Least squared problem,

$$\sup_{\|\theta\|_0 \le k} \mathbb{E}_\theta \|\theta - \hat{\theta}_{\text{MLE}}\| \lesssim k \log \frac{ep}{k}.$$

Proof. Let $h = \hat{\theta}_{\text{MLE}} - \theta$. Thus,

$$\|Z - h\|_2^2 = \|\hat{\theta}_{\text{MLE}} - y\|_2^2 \le \|\theta - y\|_2^2 = \|z\|_2^2.$$

It is equivalent to

$$\begin{aligned}\|h\|_2^2 &\le 2\langle h, z\rangle \\ &\le 2 \sup_{\|u\|_0 \le 2k} \langle u, z\rangle \\ &= 2\|h\|_2 \sup_{\|u\|_0 \le 2k, u \in S^{p-1}} \langle u, z\rangle,\end{aligned}$$

where S^{p-1} is the unit sphere in $\mathbb{R}^p$. Let $A = S^{p-1} \cap B_0(2k)$, then $\mathbb{E}\sup_{u\in A}\langle u, z\rangle \triangleq w(A)$ is the Gaussian width defined before. We have shown

$$\mathbb{E}\|h\|_2 \le 2w(A).$$

References

[Bir83] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(2):181–237, 1983.

[LC86] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York, NY, 1986.

[vdV02] Aad van der Vaart. The statistical work of Lucien Le Cam. pages 631–682, 2002.