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

Lecture 22: Functional Estimation: LeCam's Method, Risk Upper Bound

Lecturer: Yihong Wu *Scribe: Ravi Kiran Raman, Apr 19, 2016*

In the previous lecture we considered the problem of functional estimation and the idea of using LeCam's method by averaging over multiple points to obtain better estimates of the lower bound for minimax risk. In this lecture, we first use LeCam's method to obtain the lower bound and later describe an estimator obtain the matching upper bound for the minimax risk for the estimation of the ℓ_2-norm of GLM.

Consider the p-dimensional GLM. Let $\theta \in \Theta = \mathbb{R}^p$ and $X \sim \mathcal{N}(\theta, \frac{1}{n} I_p)$. Let $T(\theta) = \|\theta\|_2$. Then,

$$R^*(\Theta) = \inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^p} \mathbb{E}_\theta[(\hat{T} - T)^2] \asymp \frac{\sqrt{p}}{n}.$$

Owing to the scaling property, it suffices to prove the result for the 1-sample GLM.

22.1 LeCam's Method Lower Bound

In order to employ LeCam's method, consider the binary detection problem defined by

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \|\theta\|_2 \geq \rho \end{cases}.$$

Let $\pi(\cdot)$ be a distribution on $\{\theta : \|\theta\|_2 \geq \rho\}$, $P_0 = \mathcal{N}(0, I_p)$ and

$$P_\pi = \int \mathcal{N}(\theta, I_p)\pi(d\theta).$$

Then by LeCam's method we saw in the previous lecture that

$$R^* \geq \rho^2 \left(1 - d_{\mathrm{TV}}(P_0, P_\pi)\right) \gtrsim \rho^2,$$

when $1 - d_{\mathrm{TV}}(P_0, P_\pi) \gtrsim 0$. From the bounds on the total variational distance, we know that the above condition is satisfied when $\chi^2(P_\pi, P_0) \lesssim 1$, i.e., the χ^2 distance is bounded.

From the Ingster-Suslina method, we know that

$$\chi^2(P_\pi, P_0) = \mathbb{E}\left[G(\theta, \tilde{\theta})\right] - 1,$$

where $\theta, \tilde{\theta} \overset{i.i.d}{\sim} \pi$ and

$$G(\theta, \tilde{\theta}) = \int \frac{P_\theta(dx) P_{\tilde{\theta}}(dx)}{P_0(dx)}.$$

For the GLM,

$$G(\theta, \tilde{\theta}) = \exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right).$$

Remark 22.1. As an aside, we note that

$$\mathbb{E}\left[\exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right)\right] \geq \exp\left(\mathbb{E}\left[\left\langle \theta, \tilde{\theta} \right\rangle\right]\right) = \exp\left(\left\langle \mathbb{E}\theta, \mathbb{E}\tilde{\theta} \right\rangle\right) = \exp\left(\|\mathbb{E}\theta\|_2^2\right) > 1.$$

We now consider three priors and bound the χ^2 distance in each case.

22.1.1 Uniform distribution on sphere

Let $\theta, \tilde{\theta} \overset{i.i.d}{\sim} \text{Unif}(\rho S^{p-1})$. Let $\theta = \rho u, \tilde{\theta} = \rho \tilde{u}$ where $\|u\|_2^2 = \|\tilde{u}\|_2^2 = 1$. Hence,

$$\mathbb{E}\left[\exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right)\right] = \mathbb{E}\left[\exp\left(\rho^2 \langle u, \tilde{u} \rangle\right)\right].$$

We now exploit the fact that the inner product of directions in high dimensions is small. Let $u = \frac{Z}{\|Z\|_2}, \tilde{u} = \frac{\tilde{Z}}{\|\tilde{Z}\|_2}$, where $Z, \tilde{Z} \overset{i.i.d}{\sim} \mathcal{N}(0, I_p)$ and let $\rho^2 = c\sqrt{p}$. Then,

$$\mathbb{E}\left[\exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right)\right] = \mathbb{E}\left[\exp\left(c\sqrt{p}\frac{\left\langle Z, \tilde{Z} \right\rangle}{\|Z\|_2 \|\tilde{Z}\|_2}\right)\right] = \mathbb{E}\left[\exp\left(cY\right)\right],$$

where

$$Y = \frac{\sqrt{p}\left\langle Z, \tilde{Z} \right\rangle}{\|Z\|_2 \|\tilde{Z}\|_2}.$$

Now, by the Central Limit Theorem and the fact that $\|Z\|_2 = O_P(\sqrt{p})$,

$$\frac{\left\langle Z, \tilde{Z} \right\rangle}{\sqrt{p}} \xrightarrow{D} \mathcal{N}(0,1), \quad \frac{\|Z\|_2}{\sqrt{p}} \xrightarrow{P} 1, \quad \frac{\|\tilde{Z}\|_2}{\sqrt{p}} \xrightarrow{P} 1.$$

Thus, from Slutsky's theorem, $Y \xrightarrow{D} \mathcal{N}(0,1)$. Here we are interested in the convergence of the MGF of Y.

Remark 22.2 (Convergence of MGF)**.** Let X_n be a sequence of random variables such that $X_n \xrightarrow{D} X$. Let the tail be $T(x) \triangleq \sup_n \mathbb{P}\left[|X_n| > x\right]$. If $\forall t < \alpha$,

$$T(x)\exp\left(|t|x\right) \to 0, \quad \text{as } x \to \infty,$$

then

$$\mathbb{E}\left[\exp\left(tX_n\right)\right] \to \mathbb{E}\left[\exp\left(tX\right)\right], \forall t < \alpha.$$

Here, we have $Y \in [-\sqrt{p}, \sqrt{p}]$. Thus, by Hoeffding's inequality the tail of Y is exponentially bounded. Thus, the MGF of Y converges to the MGF of $\mathcal{N}(0,1)$ which is given by

$$\mathbb{E}\left[\exp\left(sX\right)\right] = \exp\left(\frac{1}{2}s^2\right), \quad \text{when } X \sim \mathcal{N}(0,1).$$

Thus $1 - d_{\text{TV}}(P_0, P_\pi) \gtrsim 0$ and $R^* \gtrsim \sqrt{p}$.

22.1.2 Uniform distribution on hypercube

Let $\rho = cp^{\frac{1}{4}}$ and $\theta, \tilde{\theta} \overset{\text{i.i.d}}{\sim} \text{Unif}\left(cp^{-\frac{1}{4}}\{\pm 1\}^p\right)$.

$$\mathbb{E}\left[\exp\left(\left\langle\theta,\tilde{\theta}\right\rangle\right)\right]=\mathbb{E}\left[\exp\left(\frac{c^2}{\sqrt{p}}\left\langle W,\tilde{W}\right\rangle\right)\right]=\mathbb{E}\left[\exp\left(\frac{c^2}{\sqrt{p}}G_p\right)\right],$$

where $G_p=\left\langle W,\tilde{W}\right\rangle=\sum_{i=1}^{p}W_i\tilde{W}_i$. Now,

$$\mathbb{E}\left[\exp\left(W_i\tilde{W}_i\right)\right]=\frac{1}{2}\left(\exp\left(\frac{c^2}{\sqrt{p}}\right)+\exp\left(-\frac{c^2}{\sqrt{p}}\right)\right).$$

Using Taylor's expansion, we have

$$\frac{\exp(x)+\exp(-x)}{2}=1+\frac{x^2}{2!}+\frac{x^4}{4!}+\cdots\leq\sum_n\frac{(x^2)^n}{n!}=\exp(x^2).$$

Thus we have

$$\mathbb{E}\left[\exp\left(\left\langle\theta,\tilde{\theta}\right\rangle\right)\right]=\left(\frac{1}{2}\left(\exp\left(\frac{c^2}{\sqrt{p}}\right)+\exp\left(-\frac{c^2}{\sqrt{p}}\right)\right)\right)^p\leq\exp(c^4).$$

This consequently implies that for a sufficiently small constant, the χ^2 distance is small as well. Thus, $1-d_{\mathrm{TV}}(P_0,P_\pi)\gtrsim 0$ and $R^*\gtrsim\sqrt{p}$.

22.1.3 Uniform prior on sparse vectors

Let us consider the binary hypothesis test given by

$$\begin{cases}H_0:\theta=0\\ H_1:\|\theta\|_2\geq\rho,\ \theta\in\mathbb{R}^p_+\end{cases}.$$

Now, the priors considered earlier can't be used. In this context, we shall use sparse vectors and a uniform prior to bound the χ^2 distance.

Consider the set of k-sparse vectors and let $\theta,\tilde{\theta}\overset{\text{i.i.d}}{\sim}\mathrm{Unif}\{\theta\in\{0,\epsilon\}^p:|\mathrm{supp}(\theta)|=k\}$. Let $I=\mathrm{supp}(\theta),\tilde{I}=\mathrm{supp}(\tilde{\theta})$. Let $\rho=cp^{\frac{1}{4}}$. Then

$$\|\theta\|_2=\epsilon\sqrt{k}=\rho=cp^{\frac{1}{4}}.$$

Let $k=\sqrt{p}$. Then $\epsilon=c$. Then

$$\mathbb{E}\left[\exp\left(\left\langle\theta,\tilde{\theta}\right\rangle\right)\right]=\mathbb{E}\left[\exp\left(c^2\left\langle 1_I,1_{\tilde{I}}\right\rangle\right)\right]=\mathbb{E}\left[\exp\left(c^2|\mathrm{supp}(I\cap\tilde{I})|\right)\right].$$

Owing to the symmetry of the problem, it suffices to fix I to be $\{1,\ldots,k\}$ and consider the expectation with respect to the uniform distribution on $\tilde{I}$. Thus $B=|\mathrm{supp}(I\cap\tilde{I})|$ is distributed as hypergeometric$(p,\sqrt{p},\sqrt{p})$.

Theorem 22.1 (Theorem 4, [Hoe63]). *Let the population $C=\{c_1,\ldots,c_N\}$. Let $X_1,\ldots,X_n$ denote a random sample without replacement from C and $Y_1,\ldots,Y_n$ denote a random sample with replacement. Let $f(\cdot)$ be a continuous and convex function. Then,*

$$\mathbb{E}\left[f\left(\sum_{i=1}^{n}X_i\right)\right]\leq\mathbb{E}\left[f\left(\sum_{i=1}^{n}Y_i\right)\right].$$

As a corollary of the above theorem, we have

Corollary 22.1. *Let $B \sim$ hypergeometric$(p, \sqrt{p}, \sqrt{p})$ and $\tilde{B} \sim$ binomial$(\sqrt{p}, \frac{1}{\sqrt{p}})$. Then,*

$$\mathbb{E}\left[\exp\left(sB\right)\right] \leq \mathbb{E}\left[\exp\left(s\tilde{B}\right)\right] = \left(1 - \frac{1}{\sqrt{p}} + \frac{1}{\sqrt{p}}\exp(s)\right)^{\sqrt{p}}.$$

Thus, we have,

$$\mathbb{E}\left[\exp\left(\left\langle\theta, \tilde{\theta}\right\rangle\right)\right] \leq \left(1 + \frac{1}{\sqrt{p}}\left(\exp(c^2) - 1\right)\right)^{\sqrt{p}} \leq \exp\left(\exp\left(c^2\right) - 1\right).$$

Hence for a sufficiently small c, we see that the TV distance is bounded away from 1 and thus $R^* \gtrsim \sqrt{p}$.

22.2 Risk Upper Bound

Having obtained the risk lower bound using LeCam's method, we now seek an estimator that achieves the matching upper bound on the risk. That is, given $X \sim \mathcal{N}(0, I_p)$, we seek to obtain an estimator $\hat{T} = \hat{T}(X)$ of $T = \|\theta\|_2$, such that

$$\sup_{\theta \in \mathbb{R}^p} \mathbb{E}_\theta\left[\left(\hat{T} - T\right)^2\right] \lesssim \sqrt{p}.$$

We shall first consider the plug-in estimator $\hat{T} = \|X\|_2$. Here we note from the triangle inequality that

$$|\hat{T} - T| = |\|X\|_2 - \|\theta\|_2| \leq \|Z\|_2 = O_P(\sqrt{p}).$$

Consequently, $\mathbb{E}_\theta\left[\left(\hat{T} - T\right)^2\right] \lesssim p$. However, we can verify that this bound is tight - consider the case where $\theta = 0$. This increased risk can be attributed to the presence of a bias in the estimator. That is, we have

$$\mathbb{E}_\theta\left[\|X\|_2^2\right] = \mathbb{E}_\theta\left[\|Z + \theta\|_2^2\right] = \mathbb{E}_\theta\left[\|Z\|_2^2\right] + \|\theta\|_2^2 + \mathbb{E}_\theta\left[2\left\langle Z, \theta\right\rangle\right] = p + \|\theta\|_2^2.$$

In order to negate this bias, define the estimator $\hat{T} = \sqrt{\left(\|X\|_2^2 - p\right)_+}$, where $(x)+ = \max(x, 0)$. We shall split the analysis of risk of the estimator to two cases.

Case 1: $\|\theta\|_2 \leq p^{\frac{1}{4}}$

Here we have

$$R_\theta = \mathbb{E}_\theta\left[\left(\hat{T} - \|\theta\|_2\right)^2\right] \leq 2\mathbb{E}_\theta\left[\hat{T}^2\right] + 2\|\theta\|_2^2 \leq 2\mathbb{E}\left[|S|\right] + O(\sqrt{p}),$$

where $S = \|X\|_2^2 - p$. We now note that

$$\mathbb{E}_\theta\left[\left|\|X\|_2^2 - p\right|\right] \leq \|\theta\|_2^2 + 2\mathbb{E}_\theta\left[|\left\langle\theta, Z\right\rangle|\right] + \mathbb{E}_\theta\left[\left|\|Z\|_2^2 - p\right|\right] = O_P(\sqrt{p}),$$

owing to the Central Limit Theorem and the fact that $\|\theta\|_2^2 \leq \sqrt{p}$. Using this, we have $R_\theta \lesssim \sqrt{p}, \forall \|\theta\|_2 \leq p^{\frac{1}{4}}$.

Case 2: $\|\theta\|_2 \geq p^{\frac{1}{4}}$

In this case, let us rewrite the estimation error as follows

$$\hat{T} - T = \sqrt{S_+} - \|\theta\|_2 = \frac{S_+ - \|\theta\|_2^2}{\sqrt{S_+} + \|\theta\|_2}.$$

Thus, we have

$$\begin{aligned}
|\hat{T} - T| &\leq \frac{|\left(\|X\|_2^2 - p\right)_+ - \|\theta\|_2^2|}{\|\theta\|_2} \leq \frac{|\|X\|_2^2 - p - \|\theta\|_2^2|}{\|\theta\|_2} \\
&= \frac{|\|Z\|_2^2 + \|\theta\|_2^2 + 2\langle\theta, Z\rangle - p - \|\theta\|_2^2|}{\|\theta\|_2} \\
&\leq \frac{|\|Z\|_2^2 - p|}{\|\theta\|_2} + \frac{|2\langle\theta, Z\rangle|}{\|\theta\|_2},
\end{aligned}$$

where the last step follows from the triangle inequality. Further, we have

$$|\|Z\|_2^2 - p| = O_P(\sqrt{p})$$

and

$$\frac{|2\langle\theta, Z\rangle|}{\|\theta\|_2} = |2\left\langle \frac{\theta}{\|\theta\|_2}, Z \right\rangle| = O_P(1),$$

as $\left\langle \frac{\theta}{\|\theta\|_2}, Z \right\rangle \sim \mathcal{N}(0,1)$. Thus, using the fact that $\|\theta\|_2 \geq p^{\frac{1}{4}}$, we have

$$|\hat{T} - T| \lesssim p^{\frac{1}{4}} \Leftrightarrow R_\theta \lesssim \sqrt{p}.$$

Thus, summarizing the two cases, we observe that

$$\sup_{\theta \in \mathbb{R}^p} \mathbb{E}_\theta \left[\left(\hat{T} - T \right)^2 \right] \lesssim \sqrt{p}$$

and thus $R^* \asymp \sqrt{p}$.

Example 22.1 (Covariance model and independence testing)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \Sigma)$ where Σ is a $p \times p$-dimension Covariance matrix which is to be estimated. Under this model,

- estimating Σ with $l(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_{\text{op}}$ needs $\Theta(\sqrt{p})$ samples;
- estimating $T(\Sigma) = \|\Sigma\|_{\text{op}}$ with $l(\hat{T}, T) = (\hat{T} - T)^2$ also needs $\Theta(\sqrt{p})$ samples.

Example 22.2 (Looseness of χ^2-method and sharp constant by truncated χ^2)**.** Let $X \sim \mathcal{N}(\theta, I_p)$ and $T = T(\theta) = \theta_{\max} \triangleq \max_{i \in [p]} \theta_i$. Let $l(\hat{T}, T) = \left(\hat{T} - T \right)^2$. Then,

$$R^* = \inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^p} \mathbb{E}_\theta \left[\left(\hat{T} - T \right)^2 \right] = \frac{1}{2} (1 + o(1)) \log p, \text{ as } p \to \infty.$$

The results of the above examples are proved in the next lecture.

References

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.