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

Lecture 9: Exact minimax risk for Gaussian location model, LeCam's method

Lecturer: Yihong Wu *Scribe: Siddhartha Satpathi, Feb 23, 2016 [Ed. Apr 20]*

In this lecture we consider estimation problems with no prior assumption on the structure of the parameter space. Examples of structures include sparsity, smoothness and low-rankness.

Let $X = (X_1, \ldots, X_n) \overset{i.i.d}{\sim} P_\theta$ be n samples drawn from distribution P_θ parametrized by $\theta \in \Theta$, where Θ is $\mathbb{R}^p$. Given a loss function $\ell : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$, the minimax risk is

$$R_n^*(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \ell(\theta, \hat{\theta}).$$

Two obvious observations:

- More structures lead to smaller risk. Formally, if $\Theta' \subset \Theta$, then $R_n^*(\Theta') \le R_n^*(\Theta)$.[1] Without assuming any prior structure, $\Theta = \mathbb{R}^p$, and we denote $R_n^*(\mathbb{R}^p) = R_{n,p}^*$.
- More samples lead to smaller risk. Formally, $n \mapsto R_n^*(\Theta)$ is decreasing and typically vanishing as $n \to \infty$. In the classical large-sample asymptotic regime as studied in Lecture **??**, the speed is usually "parametric", e.g., $\frac{1}{n}$ under the quadratic risk. In comparison, the focus in this course is understanding the dependency on dimension and other structural parameters without assuming large sample size. This is captured by the minimax rate. For example, we say $R_{n,p}^* \asymp \Psi_{n,p}$, when $c \le \frac{R_{n,p}^*}{\Psi_{n,p}} \le c', \ \forall\, n, p$ for some universal constants c and c'.

9.1 Log-concavity, Anderson's lemma and exact minimax risk in GLM

Definition 9.1 (Gaussian location model (GLM))**.** Let $X_1, \ldots, X_n$ be iid drawn from $\mathcal{N}(\theta, I_p)$ with $\theta \in \mathbb{R}^p$. The goal is to estimate the mean θ. Let $\hat{\theta}$ denote the estimator and $R_{n,p}^*$ denote the minimax risk under loss function $\ell(\theta, \hat{\theta})$.

Theorem 9.1. *Under GLM with quadratic loss function $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2 = \sum_{i=1}^p (\theta_i - \hat{\theta}_i)^2$, then*

$$R_{n,p}^* = \frac{p}{n}, \quad \forall n, p \in \mathbb{N}.$$

Proof. We upper bound and lower bound $R_{n,p}^*$ by $\frac{p}{n}$ in order to show equality. Let us have an estimator $\bar{X} = \frac{\sum X_i}{n} \sim \mathcal{N}(\theta, \frac{1}{n} I_p)$. Hence the risk $R_{n,p}^*$ is upper bounded by the risk obtained when using estimator $\hat{\theta} = \bar{X}$. We can compute the risk for using $\hat{\theta} = \bar{X}$ as $\frac{p}{n}$. So,

$$R_{n,p}^* \le \frac{p}{n} \tag{9.1}$$

[1] Note that this does not mean that achieving $R_n^*(\Theta')$ is computationally easier than $R_n^*(\Theta)$!

We lower bound the minimax risk $R^*_{n,p}$ by Bayes risk with prior $\pi \sim \mathcal{N}(0, sI_p)$. We can compute $R^*_\pi = \frac{sp}{sn+1}$. So,

$$R^*_{n,p} \geq R^*_\pi \overset{\lim s\to\infty}{=} \frac{p}{n} \tag{9.2}$$

Combining the upper bound and lower bound in (9.1) and (9.2), we complete the proof. □

The limitation of the above proof technique is that it only works for quadratic loss function. We next discuss a more general theorem which works over a larger range of loss functions.

Definition 9.2 (Bowl-shaped)**.** A function $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is called bowl-shaped when all its sublevel sets $K_c = \{x : \rho(x) < c\}$ for all $c \in \mathbb{R}$ are convex and symmetric (i.e. $K_c = -K_c$).

Theorem 9.2. *Consider GLM with loss functions $\ell(\theta, \hat{\theta}) = \rho(\theta - \hat{\theta})$, where $\rho : \mathbb{R}^p \to \mathbb{R}_+$ is bowl-shaped and lower-semicontinous. Then*

$$R^*_{n,p} = \mathbb{E}\rho\left(\frac{Z}{\sqrt{n}}\right),$$

where $Z \sim \mathcal{N}(0, I_p)$.

Corollary 9.1. *Let $\rho = \|.\|^q, q \geq 1$, then under GLM,*

$$R^*_{n,p} = \frac{1}{n^{q/2}}\mathbb{E}\|Z\|^q.$$

Example 9.1. Applications of Corollary 9.1:

- If $\rho = \|.\|_2^2$, then $R^*_{n,p} = \frac{1}{n}\mathbb{E}\|Z\|^2 = \frac{p}{n}$.
- If $\rho = \|.\|_\infty$, then $\mathbb{E}\|Z\|_\infty \asymp \sqrt{\log p}$ and $R^*_{n,p} \asymp \sqrt{\frac{\log p}{n}}$.
- If $\theta \in \mathbb{R}^{p\times p}$ is a matrix, and $\rho = \|.\|^2_{op} = \sigma_{\max}(\cdot)$, then $\mathbb{E}\|Z\|_{op} \asymp \sqrt{p}$ and $R^*_{n,p} \asymp \frac{p}{n}$
- If $\theta \in \mathbb{R}^{p\times p}$ is a matrix, and $\rho = \|.\|^2_F$, $R^*_{n,p} = \frac{p^2}{n}$.

Proof of Theorem 9.2. (Upper bound) Consider the estimator $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_i X_i \sim \mathcal{N}(\theta, \frac{1}{n}I_p)$. Then $\theta - \hat{=}\sqrt{\frac{1}{n}}Z$ where $Z \sim \mathcal{N}(0, I_p)$. Thus

$$R^*_{n,p} \geq \mathbb{E}[\ell(\theta, \bar{X})] = \mathbb{E}[\rho(\theta - \bar{X})] = \mathbb{E}[\rho(\frac{1}{\sqrt{n}}Z)]. \tag{9.3}$$

(Lower bound) We lower bound the minimax risk $R^*_{n,p}$ by Bayes risk R^*_π with prior $\pi = \mathcal{N}(0, sI_p)$:

$$\begin{aligned}
R^*_{n,p} &\geq R^*_\pi \\
&= \inf_{\hat{\theta}} \mathbb{E}_\pi[\rho(\theta - \hat{\theta})] \\
&= \inf_{\hat{\theta}} \mathbb{E}[\mathbb{E}[\rho(\theta - \hat{\theta})|X]] \\
&= \mathbb{E}[\inf_{\hat{\theta}} \mathbb{E}[\rho(\theta - \hat{\theta})|X]] \\
&\overset{(a)}{=} \mathbb{E}[\mathbb{E}[\rho(\theta - \mathbb{E}[\theta|X])|X]] \\
&\overset{(b)}{=} \mathbb{E}[\rho(\sqrt{\frac{s}{1+sn}}Z)] \\
&\overset{s\to\infty}{=} \lim_{s\to\infty} \mathbb{E}[\rho(\sqrt{\frac{s}{1+sn}}Z)] \\
&\overset{(c)}{=} \mathbb{E}[\lim_{s\to\infty} \rho(\sqrt{\frac{s}{1+sn}}Z)] \\
&\overset{(d)}{=} \mathbb{E}[\rho(\lim_{s\to\infty} \sqrt{\frac{s}{1+sn}}Z)] \\
&= \mathbb{E}[\rho(\frac{1}{\sqrt{n}}Z)] \qquad (9.4)
\end{aligned}$$

where (a) follows from Anderson's Lemma 9.1, (b) uses $Z \sim \mathcal{N}(0, I_p)$ or $\sqrt{\frac{s}{1+sn}}Z = (\theta - \mathbb{E}[\theta|X]) \sim \mathcal{N}(0, \frac{s}{1+sn}I_p)$ since $\theta|X \sim \mathcal{N}(\frac{sn}{1+sn}, \frac{s}{1+sn}I_p)$, (c) follows from Fatou's Lemma, and (d) follows since $\rho(\cdot)$ is a lower-semicontinuous function.

Combining the upper bound and lower bounds in (9.3) and (9.4), we can say that $R^*_{n,p} = \mathbb{E}[\rho(\frac{1}{\sqrt{n}}Z)]$. □

Lemma 9.1 (Anderson)**.** *Let $X \sim \mathcal{N}(0, \Sigma)$, and $\rho : \mathbb{R}^p \to \mathbb{R}_+$ is a bowl-shaped loss function, then*

$$\min_{y\in\mathbb{R}^p} \mathbb{E}[\rho(y + X)] = \mathbb{E}[\rho(X)].$$

In order to prove Lemma 9.1, it suffices to consider ρ being indicator functions. This is done in the next lemma, which we prove later for simpler exposition.

Lemma 9.2. *Let $K \in \mathbb{R}^p$ be a symmetric convex set and $X \sim \mathcal{N}(0, \Sigma)$ for some covariance matrix Σ. Then $\forall y \in \mathbb{R}, \mathbb{P}(X + y \in K) \leq \mathbb{P}(X \in K)$.*

Proof of Lemma 9.1. Denote the sub-level set set $K_c = \{x \in \mathbb{R}^p : \rho(x) < c\}$. Since ρ is bowl-shaped,

K_c is convex and symmetric, which satisfies the conditions of Lemma 9.2. So,

$$\begin{aligned}
\mathbb{E}[\rho(y+x)] &= \int_0^\infty \mathbb{P}(\rho(y+x) \geq c) dc, \\
&= \int_0^\infty (1 - \mathbb{P}(y+x \in K_c)) dc, \\
&\geq \int_0^\infty (1 - \mathbb{P}(x \in K_c)) dc, \\
&= \int_0^\infty \mathbb{P}(\rho(x) \geq c) dc, \\
&= \mathbb{E}[\rho(x)].
\end{aligned}$$

Hence, $\min_{y \in \mathbb{R}^p} \mathbb{E}[\rho(y+x)] = \mathbb{E}[\rho(x)]$. □

Before going into the proof of Lemma 9.2, we need the following definition.

Definition 9.3. A measure μ on $\mathbb{R}^p$ is said to be *log-concave* if

$$\mu(\lambda A + (1-\lambda)B) \geq \mu(A)^\lambda \mu(B)^{1-\lambda}$$

for all measurable $A, B \subset \mathbb{R}^p$ and any $\lambda \in [0,1]$.

The following result characterizes log-concavity of measures in terms of that of its density. See [Rin76] for a proof.

Theorem 9.3 (Prékopa)**.** *A measure μ is log-concave if and only if μ has a density f with respect to the Lebesgue measure, such that f is a log-concave function.*

Example 9.2. Examples of log-concave measures:

- Lebesgue measure: Let $\mu = \mathrm{vol}$ be the Lebesgue measure on $\mathbb{R}^p$, which satisfies Theorem 9.3 ($f(x) \equiv 1$). Then
$$\mathrm{vol}(\lambda A + (1-\lambda)B) \geq \mathrm{vol}(A)^\lambda \mathrm{vol}(B)^{1-\lambda}, \tag{9.5}$$
which implies[2] the Brunn-Minkowski inequality:
$$\mathrm{vol}(A+B)^{\frac{1}{p}} \geq \mathrm{vol}(A)^{\frac{1}{p}} + \mathrm{vol}(B)^{\frac{1}{p}}. \tag{9.6}$$
- Gaussian distribution: Let $\mu = \mathcal{N}(0, \Sigma)$, with a log-concave density f since $\log f(x) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log\det(\Sigma) - \frac{1}{2}x'\Sigma^{-1}x$ is concave.

Proof of Lemma 9.2. By Theorem 9.3, the distribution of X is log-concave. Then

$$\mathbb{P}[X \in K] \overset{(a)}{=} \mathbb{P}\Big[X \in \frac{1}{2}(K+y) + \frac{1}{2}(K-y)\Big] \tag{9.7}$$
$$\overset{(b)}{\geq} \sqrt{\mathbb{P}[X \in K-y]\mathbb{P}[X \in K+y]} \tag{9.8}$$
$$\overset{(c)}{=} \mathbb{P}[X+y \in K], \tag{9.9}$$

[2] Applying (9.5) to $A' = \mathrm{vol}(A)^{-1/p}A$, $B' = \mathrm{vol}(B)^{-1/p}B$, and $\lambda = \frac{\mathrm{vol}(A)^{1/p}}{\mathrm{vol}(A)^{1/p}+\mathrm{vol}(B)^{1/p}}$.

where (a) follows from $\frac{1}{2}(K+y)+\frac{1}{2}(K-y)=\frac{1}{2}K+\frac{1}{2}K=K$ since K is convex; (b) follows from the definition of log-concavity in Definition 9.3 with $\lambda=\frac{1}{2}$, $A=K-y=\{x-y : x\in K\}$ and $B=K+y$; (c) follows from $\mathbb{P}[X\in K+y]=\mathbb{P}[X\in -K-y]=\mathbb{P}[X+y\in K]$ since X has a symmetric distribution and K is symmetric ($K=-K$). □

9.2 LeCam's two-point argument

In this section we study a general method to obtain a lower bound on the minimax risk $R^*_{n,p}(\Theta)$.

Theorem 9.4 (LeCam's Method/two-point argument)**.** *Suppose the loss function $\ell:\Theta\times\Theta\to\mathbb{R}_+$ satisfies α-triangle inequality*

$$\ell(\theta_0,\theta_1)\le\alpha(\ell(\theta_0,\theta)+\ell(\theta_1,\theta))$$

$\forall\ \theta_0,\theta_1,\theta$ *with $\alpha>0$, then*

$$R^*_{n,p}\ge\frac{\ell(\theta_0,\theta_1)}{4\alpha}(1-d_{\mathrm{TV}}(P_{\theta_0},P_{\theta_1}))$$

Proof. In general, testing is "easier" in statistical sense than estimation. Hence, in LeCams method, we convert the estimation problem $\hat\theta^*=\arg\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbb{E}_\theta\ell(\theta,\hat\theta)$ to a hypothesis testing problem by discretizing the set Θ and obtain a lower bound on risk $R^*_{n,p}$.

For simplicity, let us break Θ into two points $\Theta'=\{\theta_1,\theta_2\}\subset\Theta$. Consider the problem, when the distribution $\mathcal{P}_\theta$ drawn from set $\{\mathcal{P}_{\theta_1},\mathcal{P}_{\theta_2}\}$. Let us consider the risk in this problem using test ψ, where,

$$\psi=\begin{cases}\theta_0 & \ell(\theta_0-\hat\theta)\le\ell(\theta_1-\hat\theta)\\ \theta_1 & \ell(\theta_1-\hat\theta)<\ell(\theta_0-\hat\theta)\end{cases}$$

for any estimate $\hat\theta$ for problem $\theta\in\Theta$

Let us denote the minimax risk obtained in this problem as $R^*(\Theta')$. Since, we are considering a simpler problem of $\theta=\theta_1$ or $\theta=\theta_2$ rather than $\theta\in\Theta$, the risk $R^*(\Theta')$ forms a lower bound to the risk $R^*_{n,p}$. So,

$$R^*_{n,p}\ge R^*(\Theta')\overset{(b)}{=}R^*_{\theta_0}\vee R^*_{\theta_1}. \tag{9.10}$$

where (b) follows from the definition of minimax risk.

Now, let $\epsilon=\ell(\theta_0,\theta_1)$. Probability of false alarm is defined $P_{\theta_0}(\psi=\theta_1)$ and probability of miss is defined as $P_{\theta_1}(\psi=\theta_0)$. Now,

$$\begin{aligned}P_{\theta_0}(\psi=\theta_1)&=P_{\theta_0}(l(\hat\theta-\theta_1)\le l(\hat\theta-\theta_0))\\&\overset{(a)}{\le}P_{\theta_0}(l(\hat\theta-\theta_0)\ge\frac{\epsilon}{2\alpha})\\&\overset{(b)}{\le}\frac{2\alpha}{\epsilon}\mathbb{E}_{\theta_0}[l(\hat\theta-\theta_0)]\end{aligned} \tag{9.11}$$

where (a) follows because event $\{l(\hat{\theta}-\theta_1) \leq l(\hat{\theta}-\theta_0)\} \subset \{l(\hat{\theta}-\theta_0) \geq \frac{\epsilon}{2\alpha}\} \subset \{l(\hat{\theta}-\theta_1) \leq \frac{\epsilon}{2\alpha}\}$. In other words, $l(\hat{\theta}-\theta_1) \leq \frac{\epsilon}{2\alpha} \implies l(\hat{\theta}-\theta_1) \leq l(\hat{\theta}-\theta_0)$. This can be verified as below,

$$\begin{aligned} \ell(\theta_0,\theta_1) &\leq \alpha(\ell(\theta_0,\theta)+\ell(\theta_1,\theta)), \\ \ell(\theta_0,\theta) \vee \ell(\theta_1,\theta) &\geq \frac{\epsilon}{2\alpha}. \end{aligned} \tag{9.12}$$

Since, $l(\hat{\theta}-\theta_1) \leq \frac{\epsilon}{2\alpha}$, together with (9.12), it implies, $l(\hat{\theta}-\theta_1) \leq \frac{\epsilon}{2\alpha} \leq \frac{\ell(\theta_0-\theta)+\ell(\theta_1-\theta)}{2}$. Now, $l(\hat{\theta}-\theta_1) \leq \frac{\ell(\theta_0-\hat{\theta})+\ell(\theta_1-\hat{\theta})}{2} \implies l(\hat{\theta}-\theta_1) \leq \ell(\theta_0-\hat{\theta})$. Therefore, we establish (a).

(b) follows from Markov's inequality.

Similarly, we can establish that the probability of miss

$$P_{\theta_1}[\psi=\theta_0] \leq \frac{2\alpha\mathbb{E}_{\theta_1}[l(\hat{\theta}-\theta_1)]}{\epsilon} \tag{9.13}$$

Now, we can say that

$$\begin{aligned} 1-TV(P_{\theta_1},P_{\theta_0}) &\leq P_{\theta_1}[\psi=\theta_0]+P_{\theta_0}(\psi=\theta_1) \\ &\overset{(a)}{\leq} \frac{2\alpha}{\epsilon}(\mathbb{E}_{\theta_1}[l(\hat{\theta}-\theta_1)]+\mathbb{E}_{\theta_0}[l(\hat{\theta}-\theta_0)]) \\ &= \frac{2\alpha}{\epsilon}(R_{\theta_0}(\hat{\theta})+R_{\theta_1}(\hat{\theta})) \\ &\leq \frac{4\alpha}{\epsilon}(R_{\theta_0}(\hat{\theta}) \vee R_{\theta_1}(\hat{\theta})) \\ &\leq \frac{4\alpha}{\epsilon}(R_{\theta_0}(\hat{\theta}) \vee R_{\theta_1}(\hat{\theta})), \end{aligned} \tag{9.14}$$

where (a) follows from (9.13) and (9.11).

Combining (9.14) with (9.10), we can say that,

$$R^*_{n,p} \geq \frac{\epsilon}{4\alpha}(1-TV(P_{\theta_1},P_{\theta_0})) \tag{9.15}$$

Since (9.15) holds for every value of θ_0,θ_1, we can write,

$$R^*_{n,p} \geq \sup_{\theta_0,\theta_1} \frac{\ell(\theta_0-\theta_1)}{4\alpha}(1-TV(P_{\theta_1},P_{\theta_0}))$$

Hence proved.

□

Example for Theorem 9.4: Suppose $\ell(\theta,\hat{\theta}) = \|\theta-\hat{\theta}\|^q, \quad q \geq 1$. Then we can easily show that $l(\cdot)$ satisfies $2^{q-1}-$triangle inequality. So, by Theorem 9.4, when $q=2$,$R^*_{n,p} \geq \sup_{\theta_0,\theta_1} \frac{1}{8}\|\theta_0-\theta_1\|^2(1-TV(P_{\theta_0},P_{\theta_1}))$.

References

[Rin76] Yosef Rinott. On convexity of measures. 4(6):1020–1026, 1976.