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

Lecture 8: Bayesian Cramér-Rao (cont'd), MLE

Lecturer: Yihong Wu *Scribe: Yanjun Li, Feb 18, 2016 [Ed. Mar 9]*

8.1 Example: Gaussian Location Model (GLM)

Let $X_i = \theta + Z_i$, where $Z_i \sim N(0,1)$, and $\theta \sim \pi = N(0,s)$. Given i.i.d. observations $\mathbf{X} = (X_1, X_2, \cdots, X_n)$, we have

$$\begin{aligned}
\chi^2(P_{\theta\mathbf{X}}||Q_{\theta\mathbf{X}}) &= \chi^2(P_{\theta\bar{X}}||Q_{\theta\bar{X}}) \\
&= \chi^2(P_\theta||Q_\theta) + \mathbb{E}_Q\left[\left(\frac{P_\theta}{Q_\theta}\right)^2 \chi^2(P_{\bar{X}|\theta}||Q_{\bar{X}|\theta})\right] \\
&= (e^{\delta^2/s} - 1) + e^{\delta^2/s}(e^{n\delta^2} - 1) \\
&= e^{\delta^2(n+\frac{1}{s})} - 1.
\end{aligned}$$

The first line follows from the fact that $\bar{X}$ is a sufficient statistic ($\theta \to \bar{X} \to \mathbf{X}$), and the information processing inequality. The second line follows from Lecture 7 (last equation, Page 5). The third line follows from

$$\chi^2\left(N(\theta,\sigma^2)||N(\theta+\delta,\sigma^2)\right) = e^{\delta^2/\sigma^2} - 1.$$

Therefore, by Bayesian HCR and Bayesian Cramér-Rao Lower Bound:

$$R^*_\pi \geq \sup_{\delta \neq 0} \frac{\delta^2}{e^{\delta^2(n+\frac{1}{s})} - 1} = \lim_{\delta \to 0} \frac{\delta^2}{e^{\delta^2(n+\frac{1}{s})} - 1} = \frac{1}{n + \frac{1}{s}} = \frac{s}{sn+1}.$$

In this case, the lower bound is tight. (It has been verified that $R^*_\pi = \frac{s}{sn+1}$.) The minimax lower bound is $R^* \geq \sup_s R^*_\pi = \frac{1}{n}$.

8.2 Classical Proof of Bayesian Cramér-Rao Lower Bound

Theorem 8.1 (Same as Theorem 7.3)**.** *If $X \sim P_\theta$, $\theta \sim \pi$, we have*

$$\mathbb{E}[(\hat{\theta}(X) - \theta)^2] \geq \frac{1}{I(\pi) + \mathbb{E}_{\theta\sim\pi}[I(\theta)]}.$$

Alternative Proof. Note that

$$\int \hat{\theta}(x)\frac{\partial}{\partial\theta}(P_\theta(x)\pi(\theta))\ \mathrm{d}\theta = 0, \tag{8.1}$$

$$\int \theta\frac{\partial}{\partial\theta}(P_\theta(x)\pi(\theta))\ \mathrm{d}\theta = -\int P_\theta(x)\pi(\theta)\ \mathrm{d}\theta, \tag{8.2}$$

where the first equation follows from the regularity condition, and the second equation follows from integration by part.

Therefore,

$$\begin{aligned}\mathbb{E}\left[(\hat{\theta}(X)-\theta)\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial\theta}\right] &= \int \mu(\mathrm{d}x)\int(\hat{\theta}(x)-\theta)\frac{\partial(P_\theta(x)\pi(\theta))}{\partial\theta}\frac{P_\theta(x)\pi(\theta)}{P_\theta(x)\pi(\theta)}\mathrm{d}\theta\\ &= \int \mu(\mathrm{d}x)\int P_\theta(x)\pi(\theta)\mathrm{d}\theta\\ &= 1,\end{aligned}$$

where the second line follows from (8.1) and (8.2).

By Cauchy-Schwarz inequality,

$$1=\mathbb{E}\left[(\hat{\theta}(X)-\theta)\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial\theta}\right]\le \mathbb{E}\left[(\hat{\theta}(X)-\theta)^2\right]\mathbb{E}\left[\left(\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial\theta}\right)^2\right].$$

Hence

$$\mathbb{E}\left[(\hat{\theta}(X)-\theta)^2\right]\ge\frac{1}{\mathbb{E}\left[\left(\frac{\partial\log P_\theta(X)}{\partial\theta}+\frac{\partial\log\pi(\theta)}{\partial\theta}\right)^2\right]}=\frac{1}{\mathbb{E}[I(\theta)]+I(\pi)}.$$

□

8.3 An Alternative Information Inequality

If we choose a uniform prior in Theorem 8.1, the resulting lower bound is zero since the Fisher information of uniform distribution is infinity. Nevertheless, it is possible to obtain an alternative information inequality involving $\mathbb{E}_{\theta\sim\text{uniform}}[I(\theta)]$; however, it should be pointed out that the lower bound applies to the minimax risk (not Bayes risk with respect to uniform prior) since the proof in act involves two prior: uniform on the interval and uniform over the two endpoints.

Theorem 8.2. *Assume the usual regularity condition:*

$$\int\frac{\partial p_\theta}{\partial x}dx=0.$$

Then

$$R^*=\inf_{\hat{\theta}}\sup_{\theta\in[\theta_0-\epsilon,\theta_0+\epsilon]}\mathbb{E}_\theta[(\theta-\hat{\theta})^2]\ge\frac{1}{(\epsilon^{-1}+\sqrt{n\bar{I}})^2}$$

where $\overline{I}$ denotes the average Fisher information:

$$\overline{I}=\frac{1}{2\epsilon}\int_{\theta_0-\epsilon}^{\theta_0+\epsilon}I(\theta)\ \mathrm{d}\theta.$$

Proof. See Problem 2 in Homework 1. □

Remark 8.1. Theorem 8.2 is a strict improvement of the closely related inequality of Chernoff-Rubin-Stein:[1]

$$\inf_{\hat{\theta}}\sup_{\theta\in[\theta_0-\epsilon,\theta_0+\epsilon]}\mathbb{E}_\theta[(\theta-\hat{\theta})^2]\ge\max_{0<\delta<1}\min\left\{\frac{\delta^2}{4},\frac{1-\epsilon}{n\bar{I}}\right\}=\frac{1}{(\epsilon^{-1}+\sqrt{n\bar{I}+1})^2}.$$

[1] This is given in [Che56, Lemma 1] without proof, which Chernoff credited to Rubin and Stein.

Both this and Theorem 8.2 suffice to prove the optimal minimax lower bound.

8.4 Maximum Likelihood Estimator (MLE)

We sketch the analysis of MLE in the classical large-sample asymptotics.

Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P_{\theta_0}$, define maximum likelihood estimator:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L_\theta(\mathbf{X}),$$

where

$$L_\theta(\mathbf{X}) = \log P_\theta^{\otimes n}(\mathbf{X}) = \sum_{i=1}^n \log P_\theta(X_i).$$

Intuition:

$$\mathbb{E}_{\theta_0}\left[L_\theta(\mathbf{X}) - L_{\theta_0}(\mathbf{X})\right] = \mathbb{E}_{\theta_0}\left[\sum_{i=1}^n \log \frac{P_\theta(X_i)}{P_{\theta_0}(X_i)}\right] = -nD(P_{\theta_0}||P_\theta) \leq 0.$$

So as long as $\theta_0 \neq \theta$, $L_\theta(\mathbf{X}) - L_{\theta_0}(\mathbf{X})$ is a random walk with negative drift. From here the consistency of MLE follows upon assuming appropriate regularity conditions.

Assuming more conditions one can obtain asymptotic normality and $\sqrt{n}$-consistency of MLE. Next, we derive a local quadratic approximation of the log-likelihood function. By Taylor expansion,

$$L_\theta(\mathbf{X}) = L_{\theta_0}(\mathbf{X}) + \sum_{i=1}^n \left.\frac{\partial \log P_\theta(X_i)}{\partial \theta}\right|_{\theta=\theta_0}(\theta - \theta_0) + \frac{1}{2}\sum_{i=1}^n \left.\frac{\partial^2 \log P_\theta(X_i)}{\partial \theta^2}\right|_{\theta=\theta_0}(\theta-\theta_0)^2 + o((\theta-\theta_0)^2). \tag{8.3}$$

Recall that

$$\mathbb{E}\left[\frac{\partial \log P_\theta(X_i)}{\partial \theta}\right] = 0, \quad \mathbb{E}\left[\left(\frac{\partial \log P_\theta(X_i)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \log P_\theta(X_i)}{\partial \theta^2}\right] = I(\theta).$$

By Central Limit Theorem,

$$\frac{1}{\sqrt{nI(\theta_0)}}\sum_{i=1}^n \frac{\partial \log P_\theta(X_i)}{\partial \theta} \xrightarrow{\text{d.}} N(0,1).$$

By (weak) Law of Large Numbers,

$$\sum_{i=1}^n \frac{\partial^2 \log P_\theta(X_i)}{\partial \theta^2} = -nI(\theta_0) + o_P(n).$$

Substituting these quantities into (8.3), we obtain a local quadratic approximation of the log-likelihood function:

$$L_\theta(\mathbf{X}) \approx L_{\theta_0}(\mathbf{X}) + \sqrt{nI(\theta_0)} \cdot Z \cdot (\theta - \theta_0) - \frac{1}{2}nI(\theta_0)(\theta-\theta_0)^2,$$

where $Z \sim N(0,1)$. Maximizing the right-hand side, we obtain:

$$\hat{\theta}_{\mathrm{MLE}} \approx \theta_0 + \frac{Z}{\sqrt{nI(\theta_0)}}.$$

Therefore, MLE achieves the locally minimax lower bound $R^* \geq \frac{1+o(1)}{nI(\theta_0)}$ (see Section 7.5 in Lecture 7).

Remark 8.2. The general asymptotic theory of MLE and achieving information bound is due to Hájek and LeCam.

8.5 Bayesian Lower Bounds for Functional Estimation

Next, we derive the Bayesian Cramér-Rao lower bound for functional estimation $\widehat{T}(X)$.

Theorem 8.3. *Let $T : \mathbb{R}^p \to \mathbb{R}$, and*

$$\begin{tikzcd} \theta \arrow[r] \arrow[d] & X \arrow[d] \\ T(\theta) & \widehat{T}(X) \end{tikzcd}$$

Then we have

$$R^*_\pi \geq (\nabla T)' I^{-1} \nabla T.$$

Proof. By similar arguments in previous lectures,

$$\chi^2(P_{\theta X}||Q_{\theta X}) \geq \chi^2(P_{T-\widehat{T}}||Q_{T-\widehat{T}}) \geq \frac{\left(\mathbb{E}_P[T-\widehat{T}] - \mathbb{E}_Q[T-\widehat{T}]\right)^2}{\mathrm{Var}_Q[T-\widehat{T}]}. \tag{8.4}$$

Let $Q(\theta) = \pi(\theta)$, and $P(\theta) = \pi(\theta - \epsilon u)$, where $u \in \mathbb{R}^p$. In order to make the marginal distribution of $P_X = Q_X$, let $P_\theta(x) = Q_{\theta - \epsilon u}(x)$. Hence the numerator and the denominator in (8.4) satisfy:

$$\begin{aligned} \left(\mathbb{E}_P[T-\widehat{T}] - \mathbb{E}_Q[T-\widehat{T}]\right)^2 &= \left(\mathbb{E}_P[T] - \mathbb{E}_Q[T]\right)^2 \\ &= \left(\int \pi(\theta) T(\theta + \epsilon u)\ \mathrm{d}\theta - \int \pi(\theta) T(\theta)\ \mathrm{d}\theta\right)^2 \\ &= \left(\int \pi(\theta) \langle \nabla T, \epsilon u \rangle + o(\epsilon)\right)^2 \\ &= \epsilon^2 \langle \mathbb{E}_\pi \nabla T, u \rangle^2 + o(\epsilon^2), \end{aligned} \tag{8.5}$$

$$\mathrm{Var}_Q[T-\widehat{T}] \leq \mathbb{E}_Q[(T-\widehat{T})^2] = R_\pi. \tag{8.6}$$

The left-hand side of (8.4) satisfies

$$\begin{aligned}
\chi^2(P_{\theta X}||Q_{\theta X}) &= \chi^2(P_\theta||Q_\theta) + \mathbb{E}_Q\left[\chi^2(P_{X|\theta}||Q_{X|\theta})\left(\frac{P_\theta}{Q_\theta}\right)^2\right] \\
&= \int \frac{(\pi(\theta-\epsilon u)-\pi(\theta))^2}{\pi(\theta)}\mathrm{d}\theta + \mathbb{E}_\pi\left[\int \frac{(Q_{\theta-\epsilon u}(x)-Q_\theta(x))^2}{Q_\theta(x)}\mathrm{d}x\left(\frac{\pi(\theta-\epsilon u)}{\pi(\theta)}\right)^2\right] \\
&= \int \frac{\epsilon^2 u'(\nabla\pi)(\nabla\pi)'u}{\pi(\theta)}\mathrm{d}\theta + \mathbb{E}_\pi\left[\int \frac{\epsilon^2 u'(\nabla_\theta Q)(\nabla_\theta Q)'u}{Q_\theta(x)}\mathrm{d}x\right] + o(\epsilon^2) \\
&= \epsilon^2 u'\left(I(\pi) + \mathbb{E}_\pi[I(\theta)]\right)u + o(\epsilon^2). \qquad (8.7)
\end{aligned}$$

Substituting (8.5), (8.6), and (8.7) into (8.4), we have

$$R^*_\pi \geq \frac{\langle \mathbb{E}_\pi \nabla T, u\rangle^2}{u'\left(I(\pi)+\mathbb{E}_\pi[I(\theta)]\right)u}$$

Locally, $\mathbb{E}_\pi \nabla T(\theta) \approx \nabla T(\theta_0)$, and $I(\pi) + \mathbb{E}_\pi[I(\theta)] \approx I(\theta_0)$. Hence

$$\boxed{R^*_\pi \geq \sup_u \frac{\langle \nabla T(\theta_0), u\rangle^2}{u' I(\theta_0) u} = (\nabla T(\theta_0))' I^{-1}(\theta_0)\nabla T(\theta_0).}$$

The maximum is attained when $u = I^{-1}(\theta_0)\nabla T(\theta_0)$.[2] □

Remark 8.3. The maximum likelihood estimator satisfies $T(\hat\theta_{\mathrm{MLE}}) = T(\theta_0 + \frac{1}{\sqrt{n}}Z)$, where $Z \sim N(0, I^{-1}(\theta_0))$. Hence

$$T(\hat\theta_{\mathrm{MLE}}) \sim N\left(T(\theta_0), \frac{1}{n}(\nabla T(\theta_0))' I^{-1}(\theta_0)(\nabla T(\theta_0))\right).$$

The maximum likelihood estimator again asymptotically achieves the locally minimax lower bound.

8.6 Example: Classical asymptotics of entropy estimation

Corollary 8.1. *Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} p \in \mathcal{M}_k$, where $\mathcal{M}_k$ denotes the set of probability distributions over $[k] = \{1, \ldots, k\}$. Then the minimax quadratic risk of entropy estimation satisfies*

$$R^* = \inf_{\widehat{H}} \sup_{P \in \mathcal{M}_k} \mathbb{E}[(\widehat{H} - H)^2] = \frac{1}{n}\left(\max_{p \in \mathcal{M}_k} V(p) + o(1)\right), \quad n \to \infty$$

where

$$\begin{aligned}
H(p) &= \sum_{i=1}^k p_i \log\frac{1}{p_i} = \mathbb{E}\left[\log\frac{1}{p(X)}\right], \\
V(p) &= \mathrm{Var}\left(\log\frac{1}{p(X)}\right)
\end{aligned}$$

[2] This can be shown, for example, by letting $\tilde{u} = I^{-\frac{1}{2}}(\theta_0)u$.

Note: $\max_{p\in\mathcal{M}_k} V(p) \le \log^2 k$ for all $k \ge 3$ (see [PPV10, (464)]).

Proof. We have $H : \Theta \to \mathbb{R}^+$, where $\theta = (p_1, p_2, \cdots, p_{k-1})$.[3] Therefore,

$$\frac{\partial H}{\partial p_i} = \log \frac{p_k}{p_i}, \quad i = 1, 2, \cdots, k-1.$$

Next, we compute the Fisher Information matrix:

$$I(\theta)_{ij} = -\mathbb{E}\left[\frac{\partial^2 \log p(X)}{\partial p_i \partial p_j}\right] = \begin{cases} \frac{1}{p_i} + \frac{1}{p_k} & \text{if } i = j \\ \frac{1}{p_k} & \text{if } i \ne j \end{cases}.$$

Therefore,

$$I(\theta) = \begin{bmatrix} \frac{1}{p_1} & & \\ & \ddots & \\ & & \frac{1}{p_{k-1}} \end{bmatrix} + \frac{1}{p_k}\mathbf{1}\mathbf{1}'.$$

By Matrix Inversion Lemma,[4] we have

$$I^{-1}(\theta) = \begin{bmatrix} p_1 & & \\ & \ddots & \\ & & p_{k-1} \end{bmatrix} + \begin{bmatrix} p_1 \\ \vdots \\ p_{k-1} \end{bmatrix} \begin{bmatrix} p_1 & \cdots & p_{k-1} \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \nabla H' I^{-1}(\theta) \nabla H &= \sum_{i=1}^{k-1} p_i \log^2 \frac{p_k}{p_i} - \left(\sum_{i=1}^{k-1} p_i \log \frac{p_k}{p_i}\right)^2 \\ &= \sum_{i=1}^{k} p_i \log^2 \frac{1}{p_i} + \log^2 \frac{1}{p_k} - 2\sum_{i=1}^{k} p_i \log \frac{1}{p_i} \log \frac{1}{p_k} - \left(\left(\sum_{i=1}^{k} p_i \log \frac{1}{p_i}\right) - \log \frac{1}{p_k}\right)^2 \\ &= \sum_{i=1}^{k} p_i \log^2 \frac{1}{p_i} - \left(\sum_{i=1}^{k} p_i \log \frac{1}{p_i}\right)^2 \\ &= \mathbb{E}\left[\log^2 \frac{1}{p(X)}\right] - \left(\mathbb{E}\left[\log \frac{1}{p(X)}\right]\right)^2 = \mathrm{Var}\left[\log \frac{1}{p(X)}\right] = V(p). \end{aligned}$$

Given n samples, the Fisher Information matrix is $nI(\theta)$. By Theorem 8.3,

$$R^* \ge \frac{1+o(1)}{n} \nabla H' I^{-1}(\theta) \nabla H = \frac{1+o(1)}{n} V(p). \qquad \square$$

References

[Che56] Herman Chernoff. Large-sample theory: Parametric case. *The Annals of Mathematical Statistics*, 27(1):1–22, 1956.

[PPV10] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 56(5):2307–2359, May 2010.

[3] $p_k = 1 - p_1 - \cdots - p_{k-1}$.

[4] $(A + UCV)^{-1} = A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$.