Feynman Liang
STAT 210A
Lecture #7
9/13/2016

1 Review: UMVU Estimators

$T(X)$ complete sufficient:

$$\implies \exists \text{at most 1 unbiased } \delta(T(X)) \tag{1.1}$$
$$\implies \text{That one } \textit{best} \text{ for } \textit{any} \text{ convex loss, } \forall\theta \tag{1.2}$$

This givese us a strategy for coming up with UMVUs. Can find any unbiased estimator that is only a function of T, or can Rao-Blackwellize any unbiased estimator.

2 Log-likelihood and Score

Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ for $\Theta \subset \mathbb{R}^d$. For today (not necessary), assume *common support* i.e. $\mathcal{X} = \{x : p_\theta(x) > 0\}$ same $\forall\theta$.

Definition 2.1 (Log-likelihood function)**.**

$$l(\theta; x) = \log p_\theta(x) \tag{2.1}$$

Definition 2.2 (Score function)**.** If $l(\theta, x)$ is differentiable, the *score function*

$$\nabla l(\theta; x) \tag{2.2}$$

Assuming $\mathcal{P}$ is nice enough to differentiate under the integral, some useful facts:

- $1 = \int_{\mathcal{X}} e^{l(\theta;x)} d\mu(x)$
- $0 = \int_{\mathcal{X}} \frac{\partial}{\partial\theta_j} l(\theta; x) e^{l(\theta;x)} d\mu(x)$
- $\mathbb{E}_\theta[\nabla l(\theta; x)] = 0$
- $$0 = \int_{\mathcal{X}} \left[\frac{\partial^2}{\partial\theta_j \partial\theta_k} l + \frac{\partial}{\partial\theta_j} l \frac{\partial}{\partial\theta_k} l \right] e^l d\mu \tag{2.3}$$
  $$= \mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta_j \partial\theta_k} l \right] + \operatorname{Cov}_\theta \left(\frac{\partial}{\partial\theta_j} l, \frac{\partial}{\partial\theta_k} l \right) \tag{2.4}$$
- $\operatorname{Var}_\theta(\nabla l(\theta; x)) = \mathbb{E}_\theta \left[-\nabla^2 l(\theta; x) \right]$

Definition 2.3 (Fisher Information)**.**

$$J(\theta) = \mathbb{E}_\theta\left[-\nabla^2 l(\theta; x)\right] = \text{Var}_\theta(\nabla l(\theta)) \tag{2.5}$$

provided $l(\theta; x) \in C^2(\Theta)$

Suppose $\delta(X)$ is an *unbiased* estimator for $g(\theta)$

$$\begin{aligned}
\mathbb{E}_\theta[\delta(X)] &= g(\theta) &(2.6)\\
g(\theta) &= \int_{\mathcal{X}} \delta(x) e^{l(\theta)} d\mu(x) &(2.7)\\
\nabla g(\theta) &= \int \delta \nabla l(\theta) e^{l(\theta)} d\mu &(2.8)\\
&= \text{Cov}_\theta(\delta, \nabla l(\theta)) &(2.9)
\end{aligned}$$

Theorem 2.4 (Information bound a.k.a. Cramer-Rao lower bound (CRLB))**.** *In the* 1-*parameter case i.e.* $\theta \in \mathbb{R}$

$$\begin{aligned}
&Var_\theta(\delta) Var_\theta(l'(\theta)) \geq Cov_\theta(\delta, l')^2 &(2.10)\\
\implies\ &Var_\theta(\delta) \geq \frac{g'(\theta)^2}{J(\theta)} &(2.11)
\end{aligned}$$

For multiple parameters:

$$Var_\theta(\delta) \geq (\nabla g(\theta))'[J(\theta)^{-1}]\nabla g(\theta) \tag{2.12}$$

Example 2.5 (iid samples)**.** $X_1, \cdots, X_n \overset{\text{iid}}{\sim} p_\theta^{(1)}(x)$ for $\theta \in \Theta$.

$$\begin{aligned}
l(\theta; x) &= \sum_{i=1}^{n} l_i(\theta; x_i) &(2.13)\\
J(\theta) &= \text{Var}_\theta(\nabla l(\theta; x)) &(2.14)\\
&= nJ_1(\theta) &(2.15)
\end{aligned}$$

This shows that with n i.i.d. samples, we have n times more information than the information from a single sample $J_1(\theta)$.

Corollary 2.6. *With n i.i.d. samples*

$$Var_\theta(\delta) \geq \frac{g'(\theta)^2}{J(\theta)} \asymp \frac{1}{n} \tag{2.16}$$

Definition 2.7. $f(n) \asymp g(n)$ means

$$0 < \liminf_n \frac{f(n)}{g(n)} \leq \limsup_n \frac{f(n)}{g(n)} < \infty \tag{2.17}$$

CRLB is not necessarily attainable, but

Definition 2.8. $\delta(X)$ is *efficient* if $\text{Var}_\theta(\delta) = \text{CRLB}$.
If $\frac{\text{CRLB}}{\text{Var}_\theta(\delta)} = 0.7$, we say *70% efficient*.

We can write $\frac{\text{CRLB}}{\text{Var}_\theta(\delta)} = \text{Corr}^2_\theta(\delta, \nabla l)$ so the score function is in some sense playing the role of a local sufficient statistic and we would like an estimator δ to be more correlated to ∇l.

$$\text{Corr}^2_\theta(\delta, \nabla l) = 1 \iff \delta \text{ efficient} \tag{2.18}$$

2.1 Exponential Families

$$p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x) \tag{2.19}$$
$$l(\eta; x) = \eta' T(x) - A(\eta) + \log h(x) \tag{2.20}$$
$$\nabla l(\eta; x) = T(x) - \nabla A(\eta) \tag{2.21}$$
$$= T(x) - \mathbb{E}_\eta[T(x)] \tag{2.22}$$

The score $\nabla l(\eta; x)$ is equal to $T(x)$ up to a constant offset term $\mathbb{E}_\eta(T(X))$ which makes $\mathbb{E}_\theta \nabla l(\eta; x) = 0$.

$$\text{Var}_\eta(\nabla l(\eta)) = \text{Var}_\eta(T(x)) = \underbrace{\nabla^2 A(\eta)}_{\text{not random}} = -\nabla^2 l(\eta; x) = \mathbb{E}_\eta[-\nabla^2 l(\eta; x)] \tag{2.23}$$

So all are equal to the Fisher information for exponential families, and the Fisher information depends only on η i.e. is independent of x.

2.2 Relaxing regularity assumptions on $l(\theta; x)$

CRLB requires differentiation of e^l under integral. Instead, can consider a finite-difference version for the score. For some finite amount ϵ

$$L(x) - 1 = \frac{p_{\theta+\epsilon}(x)}{p_\theta(x)} - 1 = e^{l(\theta+\epsilon;x) - l(\theta;x)} - 1 \tag{2.24}$$
$$\approx \epsilon' \nabla l(\theta; x) \tag{2.25}$$

$L(x)$ is the *likelihood ratio*.

$$\mathbb{E}_\theta \left[L(x) - 1 \right] = \int_{\mathcal{X}} (p_{\theta+\epsilon}(x)/p_\theta(x) - 1) p_\theta d\mu \tag{2.26}$$
$$= \int_{\mathcal{X}} (p_{\theta+\epsilon}(x) - p_\theta(x)) d\mu = 1 - 1 = 0 \tag{2.27}$$
$$\tag{2.28}$$

$$\mathrm{Cov}_\theta[\delta, L(x)-1] = \int_{\mathcal{X}} \delta(p_{\theta+\epsilon}/p_\theta - 1)p_\theta d\mu \tag{2.29}$$
$$= g(\theta+\epsilon) - g(\theta) \tag{2.30}$$

Theorem 2.9 (Hammersley-Chapman-Robin (H-C-R))**.** *The above facts imply*

$$Var_\theta(\delta) \geq \frac{(g(\theta+\epsilon)-g(\theta))^2}{\mathbb{E}_\theta[(L(x)-1)]^2} \tag{2.31}$$

The previous CRLB can be viewed as the infinitismal case of this, where we multiply the numberator and denominator by $1/\epsilon^2$ and tale $\epsilon \to 0$.

3 Bayes risk minimization

A problem with estimators is that some are better than others depending on choice of $\theta \in \Theta$.

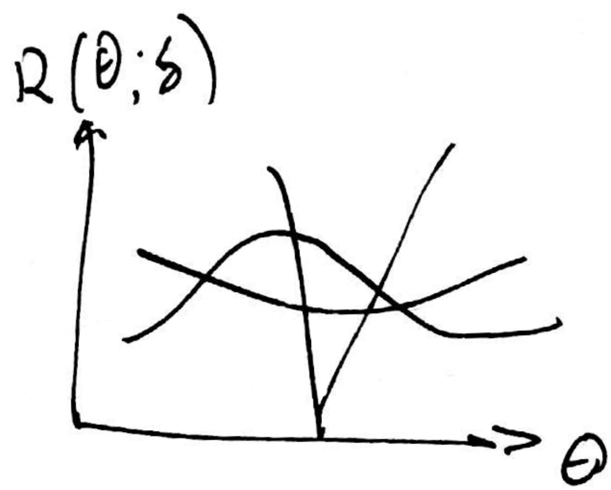

Figure 1: Different estimators have different risks $R(\theta, \delta)$ depending on choice of θ

Suppose we weight our parameter space with a weight function $w(\theta)$. Then the Bayes risk

$$\int R(\theta;\delta)w(\theta)d\theta = \mathbb{E}[R(\theta,\delta)] \tag{3.1}$$

where $\theta \sim \frac{w(\theta)}{\int_\Omega w(\theta)d\theta}$.