UC Berkeley
Department of Statistics
Fall 2016

STAT 210A: Theoretical Statistics

Problem Set 1

Student: Ivana Malenica **Due:** Thursday, Sep. 8

Note: All measure-theoretic niceties about conditioning on measure-zero sets, almost-sure equality vs. actual equality, "all functions" vs. "all measurable functions," etc. are disregarded for the time being.

1. Risk of a shrinkage estimator

Let θ denote the proportion of people working in a company who are left-handed, and suppose we are in charge of ordering equipment and need to estimate θ. Let X denote the number of left-handers in a sample of size n from the company (for simplicity, assume we sample with replacement).

It is known that 10% of the U.S. population is left-handed. Instead of using the "obvious" estimator $\hat{\theta}_0(X) = X/n$, we could "shrink" the estimator toward 10% by using:

$$\hat{\theta}_1(X) \quad = \quad 0.2 \cdot 10\% + 0.8 \cdot \frac{X}{n},$$

Let $\text{MSE}_n(\theta, \hat{\theta})$ denote the mean squared error of an estimator $\hat{\theta}$, as a function of the sample size n and true parameter θ.

(a) Find $\text{MSE}_n(\theta, \hat{\theta}_i)$ for $i = 0, 1$.

Solution:

Note that X Bin(n,θ). Then we have that:

$$\begin{aligned} MSE_n(\theta, \hat{\theta}_0) &= E(\hat{\theta}_0 - \theta)^2 \\ &= Var(\hat{\theta}_0) + Bias(\hat{\theta}_0)^2 \\ &= \frac{1}{n^2} Var_\theta(X) + (\frac{1}{n} E_\theta(X) - \theta)^2 \\ &= \frac{1}{n}\theta(1-\theta) \end{aligned}$$

$$\begin{aligned} MSE_n(\theta, \hat{\theta}_1) &= E(\hat{\theta}_1 - \theta)^2 \\ &= Var(\hat{\theta}_1) + Bias(\hat{\theta}_1)^2 \\ &= (\frac{0.8}{n})^2 Var_\theta(X) + (0.02 + \frac{0.8}{n} E_\theta(X) - \theta)^2 \\ &= \frac{0.64}{n}\theta(1-\theta) + (0.2(0.1-\theta))^2 \end{aligned}$$

(b) For what values of θ is

$$\frac{\text{MSE}_n(\theta, \hat{\theta}_1)}{\text{MSE}_n(\theta, \hat{\theta}_0)} < 1?$$

Give the answer as a function of n. What happens as $n \to \infty$?

Solution:

We need $MSE_n(\theta, \hat{\theta_1}) < MSE_n(\theta, \hat{\theta_0})$. This is possible for:

$$\frac{0.64}{n}\theta(1-\theta) + (0.2(0.1-\theta))^2 < \frac{1}{n}\theta(1-\theta) => (n+9)\theta^2 - (0.2n+9)\theta + 0.01n < 0$$

which solves for: $\theta \in \frac{(0.2n+9)\pm\sqrt{(0.2n+9)^2-4(n+9)(0.01n)}}{2(n+9)} = \frac{n+45\pm 9\sqrt{n+45}}{10n+90}$

Note that $\lim_{n\to\infty} \frac{n+45\pm 9\sqrt{n+45}}{10n+90} \to \{0.1\}$.

2. Convexity of $A(\eta)$ and Ξ

Let $\mathcal{P} = \{p_\eta : \eta \in \Xi\}$ denote an s-parameter exponential family in canonical form

$$p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x), \qquad A(\eta) = \log \int_{\mathcal{X}} e^{\eta' T(x)} h(x)\, d\mu(x),$$

where $\Xi = \{\eta : A(\eta) < \infty\}$ is the natural parameter space.

Recall Hölder's inequality: if $q_1, q_2 \geq 1$ with $q_1^{-1} + q_2^{-1} = 1$, and f_1 and f_2 are (μ-measurable) functions from $\mathcal{X}$ to $\mathbb{R}$, then

$$\|f_1 f_2\|_{L^1(\mu)} \leq \|f_1\|_{L^{q_1}(\mu)} \|f_2\|_{L^{q_2}(\mu)}, \quad \text{where } \|f\|_{L^q(\mu)} = \left(\int_{\mathcal{X}} |f(x)|^q\, d\mu(x)\right)^{1/q}.$$

(Note that $q_1 = q_2 = 2$ reduces to Cauchy-Schwarz).

(a) Show that $A(\eta) : \mathbb{R}^s \to [0, \infty]$ is a convex function: that is, for *any* $\eta_1, \eta_2 \in \mathbb{R}^s$ (not just in Ξ), and $c \in [0,1]$ then

$$A(c\eta_1 + (1-c)\eta_2) \leq cA(\eta_1) + (1-c)A(\eta_2)$$

Solution:

Let $\eta = c\eta_1 + (1-c)\eta_2$ with $c \in [0,1]$ and $\eta_1, \eta_2 \in \mathbb{R}^s$. Then $\frac{1}{c} \geqslant 1$ and $\frac{1}{1-c} \geqslant 1$ and Hölder's inequality with $q_1 = \frac{1}{c}$ and $q_2 = \frac{1}{1-c}$ implies:

$$\begin{aligned}
A(\eta) &= \log \int_{\mathcal{X}} e^{\eta' T(x)} h(x)\, d\mu(x) \\
&= \log[(\int_{\mathcal{X}} (e^{\eta_1' T(x)})^c h(x)\, d\mu(x))(\int_{\mathcal{X}} (e^{\eta_2' T(x)})^{1-c} h(x)\, d\mu(x))] \\
&\leqslant \log[(\int_{\mathcal{X}} (e^{\eta_1' T(x)})^{\frac{c}{c}} h(x)\, d\mu(x))^c (\int_{\mathcal{X}} (e^{\eta_2' T(x)})^{\frac{1-c}{1-c}} h(x)\, d\mu(x))^{1-c}] \\
&= c\log(\int_{\mathcal{X}} e^{\eta_1' T(x)} h(x)\, d\mu(x)) + (1-c)\log(\int_{\mathcal{X}} e^{\eta_2' T(x)} h(x)\, d\mu(x)) \\
&= cA(\eta_1) + (1-c)A(\eta_2)
\end{aligned}$$

(b) Conclude that $\Xi \subset \mathbb{R}^s$ is convex.

Solution:

Let η_1 and η_2 be in the natural parameter space and $0 < c < 1$. Thus, $A(\eta_1)$ and $A(\eta_2)$ are finite. We need to show that $\eta = c\eta_1 + (1-c)\eta_2$ belongs to Ξ, so that $A(\eta) < \infty$. By part (a) we know that $A(\eta) < \infty$ since $A(\eta_1)$ and $A(\eta_2)$ are finite and $0 < c < 1$. Therefore, $\eta \in \Xi$.

3. Expectation of an increasing function

(a) Assume $X \sim P$ is a real-valued random variable. Show that if $f(x)$ and $g(x)$ are non-decreasing functions of x, then

$$Cov(f(X), g(X)) \geqslant 0$$

<u>**Solution:**</u>

Let X_1,X_2 $\sim$ P. Since X_1 and X_2 are iid, we have that:

$$cov(f(X_1), g(X_2)) = cov(f(X_2), g(X_1)) = 0$$

Note that for non-decreasing functions f and g and X_1,X_2 $\sim$ P we have that:

$$cov(f(X_1), g(X_1)) = cov(f(X_2), g(X_2))$$

Then, cov(f(X_1), g(X_1)) + cov(f(X_2), g(X_2)) = 2cov(f(X_1), g(X_1)) and by independence we have that 2cov$(f(X_1), g(X_1))$ = cov$(f(X_1) - f(X_2), g(X_1) - g(X_2))$ equals:

$$\begin{aligned} &\mathbb{E}(f(X_1)g(X_1)) - \mathbb{E}(f(X_1)g(X_2)) - \mathbb{E}(f(X_2)g(X_1)) \\ &\quad + \mathbb{E}(f(X_2)g(X_2)) - \mathbb{E}(f(X_1))\mathbb{E}(g(X_1)) + \mathbb{E}(f(X_1))\mathbb{E}(g(X_2)) \\ &\quad + \mathbb{E}(f(X_2))\mathbb{E}(g(X_1)) - \mathbb{E}(f(X_2))\mathbb{E}(g(X_2)) \end{aligned} \tag{1}$$

For non-decreasing functions f and g, $f(X_1) - f(X_2)$ and $g(X_1) - g(X_2)$ carry the same sign as $X_1 - X_2$ and $(f(X_1) - f(X_2))(g(X_1) - g(X_2)) \geqslant 0$. Therefore, cov$(f(X_1) - f(X_2), g(X_1) - g(X_2))$ will be positive and hence 2cov$(f(X_1), g(X_1))$ as well.

(b) Let $p_\eta(x)$ be a one parameter canonical exponential family generated by $T(x) = x$ and $h(x)$, where $x \in \mathcal{X} \subset \mathbb{R}$, i.e.

$$p_\eta(x) = e^{\eta x - A(\eta)} h(x).$$

Let $\psi(x)$ be any non-decreasing function. Show that, if $\eta \in \Xi^o$, $E_\eta(\psi(X))$ is non-decreasing in η.

<u>**Solution:**</u>

Let $\eta \in \Xi^o$. We want to show $\frac{\delta}{\delta\eta}\mathbb{E}_\eta[\psi(x)] \geqslant 0$. If ψ and $\frac{\delta}{\delta\eta}$ are continuous, and using part (a) we have that:

$$\begin{aligned} \frac{\delta}{\delta\eta}\mathbb{E}_\eta[\psi(x)] &= \frac{\delta}{\delta\eta}\int_{\mathcal{X}} \psi(x)p_\eta(x)d\mu(x) \\ &= \int_{\mathcal{X}} \psi(x)\frac{\delta}{\delta\eta}e^{\eta x - A(\eta)}h(x)d\mu(x) \\ &= \int_{\mathcal{X}} \psi(x)xe^{\eta x - A(\eta)}h(x)d\mu(x) - \int_{\mathcal{X}} \psi(x)\mathbb{E}_\eta[x]e^{\eta x - A(\eta)}h(x)d\mu(x) \\ &= \mathbb{E}_\eta[x\psi(x)] - \mathbb{E}_\eta[x]\mathbb{E}_\eta[\psi(x)] \\ &\geqslant 0 \end{aligned}$$

4. Exponential families maximize entropy

The entropy (with respect to μ) of a random variable X with density p, is defined by

$$h(p) = \mathbb{E}_p(-\log p(X)) = -\int_{p(x)>0} \log(p(x))p(x)\, d\mu(x)$$

This quantity arises naturally in information theory as a minimal expected code length. Now consider the problem of maximizing $h(p)$ subject to the constraints that p is a probability density with $\mathbb{E}_p[T(X)] = \alpha$, for some $\alpha \in \mathbb{R}^s$, $T:\ \mathcal{X} \to \mathbb{R}^s$. That is, $p(x) > 0$, $\int p(x)d\mu(x) = 1$, and $\int p(x)T_j(x)\, d\mu(x) = \alpha_j$, for $1 \leq j \leq s$

(a) Assume $\mathcal{X}$ is a finite set and α is in the convex hull of $T(\mathcal{X}) = \{T(x) :\ x \in \mathcal{X}\}$. Show that the optimal p^* is in the s-parameter exponential family

$$p_\eta(x) = e^{\eta' T(x) - A(\eta)},$$

with parameter $\eta^* \in \mathbb{R}^s$ chosen so that p_{η^*} satisfies the constraints.

Solution:

We can set up the following maximization problem:
maximize $h(p) = \mathbb{E}_p(-\log p(X)) = -\int_{p(x)>0} \log(p(x))p(x)\, d\mu(x)$
subject to $p(x) > 0$, $\int p(x)d\mu(x) = 1$, and $\int p(x)T_j(x)\, d\mu(x) = \alpha_j$, for $1 \leq j \leq s$

We introduce the following Lagrange multipliers:
$\eta(x)$ for the constraint $p(x) > 0$
η_0 for the constraint $\int p(x)d\mu(x) = 1$
η_j for the constraint $\int p(x)T_j(x)\, d\mu(x) = \alpha_j$, for $1 \leq j \leq s$

Note that the Lagrangian is:

$$\begin{aligned}\varphi(p, \eta(x), \eta_0, \eta_j) = &-\int_{\mathcal{X}} \log(p(x))p(x)\, d\mu(x) \\ &+ \int_{\mathcal{X}} \eta(x)p(x)\, d\mu(x) + \eta_0(\int_{\mathcal{X}} p(x)\, d\mu(x) - 1) \\ &+ \sum_j \eta_j(\int_{\mathcal{X}} T_j(x)p(x)\, d\mu(x) - \alpha_j)\end{aligned} \tag{2}$$

Since $\mathcal{X}$ is finite, we obtain the following:

$$\begin{aligned}0 &= \frac{\delta\varphi(p, \eta(x), \eta_0, \eta_j)}{\delta p(x)} \\ &= -\log p(x) - 1 + \eta(x) + \eta_0 + \sum_j T_j(x)\eta_j\end{aligned} \tag{3}$$

Note that $p(x) > 0$ is an unnecessary constraint from this setup. Since the optimization is over a compact set which is nonempty as α is in the convex hull of T($\mathcal{X}$), both η_0 and η_j can be picked such that the 2 constraints are satisfied. The optimal p is:

$$\log p(x) = \sum_j T_j(x)\eta_j + \eta_0 - 1 \Rightarrow p(x) = e^{\sum_j T_j(x)\eta_j + \eta_0 - 1}$$

With a bit of algebra manipulation, we have that for $\eta_0 = 1 - \log \int e^{\eta T(x)} h(x)$, $p(x) = e^{<\eta, T(x)> - A(\eta)}$.

(b) Blithely applying the result of (a) to non-finite $\mathcal{X}$, find the distribution that maximizes the entropy (with respect to the Lebesgue measure), subject to the constraint that $\mathbb{E}(X) = \mu, \text{Var}(X) = \sigma^2$.

Solution:

We can set up the following maximization problem:
maximize $h(p) = \mathbb{E}_p(-\log p(X)) = -\int_{p(x)>0} \log(p(x))p(x)\, d\mu(x)$
subject to $E(X) = \mu$, $Var(X) = \sigma^2$, $\int p(x)d\mu(x) = 1$

We introduce the following Lagrange multipliers:
λ_0 for the constraint $\int p(x)d\mu(x) = 1$
λ for the constraint $\int p(x)(x-\mu)^2 d\mu(x) = 1$

Note that the Lagrangian is:

$$\begin{aligned}\varphi(p, \lambda_0, \lambda) = &-\int_{\mathcal{X}} \log(p(x))p(x)\, d\mu(x) + \lambda_0(\int_{\mathcal{X}} p(x)\, d\mu(x) - 1) \\ &+ \lambda(\int_{\mathcal{X}} (x-\mu)^2 p(x)\, d\mu(x) - \sigma^2)\end{aligned} \tag{4}$$

Thus:

$$0 = \frac{\delta\varphi(p, \lambda_0, \lambda)}{\delta p(x)} = -\log p(x) - 1 + \lambda_0 + \lambda$$

The optimal p is:

$$\log p(x) = \lambda(x-\mu)^2 + \lambda_0 - 1 \Rightarrow p(x) = e^{\lambda(x-\mu)^2+\lambda_0-1}$$

Using part (a), the distribution should be of the $e^{\lambda_1 x + \lambda_2 x^2 + A(\lambda)}$ form. We recognize this as the exponential family form of the normal distribution with parameters μ and σ^2.

5. Minimal sufficiency of the likelihood ratio

Suppose that $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ is a family of densities defined with respect to a common measure μ on $\mathcal{X}$. Assume $\Theta = \{\theta_1, \ldots, \theta_m\}$ is a finite set, and there exists some $\theta \in \Theta$ such that $p_\theta(x) > 0, \forall x \in \mathcal{X}$ (without loss of generality we can assume it is θ_1).

(a) Prove that the *likelihood ratio*, defined as

$$T(X) = \left(\frac{p_\theta(X)}{p_{\theta_1}(X)}\right)_{\theta\in\Theta}$$

is minimal sufficient. (Note: $\mathcal{X}$ is not necessarily finite).

Solution:

Note that $\Theta = \{\theta_1, ...\theta_m\}$ is a finite set, thus we can represent $T(x)$ as a vector of $m-1$ likelihood ratios:

$$T(x) = [\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} ... \frac{p_{\theta_m}(x)}{p_{\theta_1}(x)}]$$

Define a unit vector e_i. Then, we have $p_\theta(x) = e^t T(x) p_{\theta_1}(x)$. By applying the factorization theorem with $g_\theta(T(x)) = e^t T(x)$ and $h(x) = p_{\theta_1}(x)$ we have that the likelihood ratio is sufficient.

If $p_\theta(x) \propto_\theta p_\theta(y)$, $\exists$ c(x,y) such that $\forall\theta$ $p_\theta(x) = c(x,y)p_\theta(y)$. Thus,

$$\frac{p_\theta(x)}{p_{\theta_1}(x)} = \frac{p_\theta(y)}{p_{\theta_1}(y)}$$

and the likelihood ratio is minimal.

(b) Show by counterexample that the *likelihood function*, defined as

$$T(X) = (p_\theta(X))_{\theta\in\Theta}$$

is *not*, in general, minimal sufficient.

Solution:

Let X_1 and X_2 be independent, with $X_1 \sim \text{Bern}(\theta)$ and $X_2 \sim \text{Bern}(\frac{2}{3})$. Then:

$$p_\theta(X_1, X2) = \theta^{X_1}(1-\theta)^{1-X_1}\frac{2^{X_2}}{3}$$

By the factorization theorem, X_1 is clearly sufficient. However, $p_\theta(X_1, X2)$ cannot be written as a function of X_1 only (notice that $p_\theta(X_1, 1) = 2p_\theta(X_1, 0)$).