Statistical Theory
Quarter 3

James Pringle
taught by Constantine the Greek

22 January 2013 - 15 March 2013

1 Tuesday, 22 January 2013

Minimizing loss. Followed the online slides and packet.

2 Thursday, 24 January 2013

View the slides for chapter 1.

3 Tuesday, 29 January 2013

If s_n is Bayes with respect to π_n with $L(\pi_n, s_n) \to C$, and if there exists a strategy so $L(\theta, s_0) \leq C$ for all θ, then

$$\inf_s \sup_\theta L(\theta, s) = C = \sup_\theta L(\theta, s_0)$$

i.e. s_0 is a minimax.

We are putting a foundation of what statistics is by looking at states of nature that are known to us. We are also looking at our possible observations in relation to those states of nature. Then there is a loss function combining them all. The class of admissable strategies is essentially the same as the class of Bayes estimators.

Let I be an indicator function. I_i is the indicator for the i-th θ. By design, we have that $P(I_i = 1|E_{13}, E_{12} = P(I_i = 1|E_{12})$.

Later, we have $P(E_{12,i}, I_i, E^{obs}_{13,i}|\theta) = \int P(E_{12,i}, I_i, E_{-13,1}|\theta) dE^{miss}_{12,i}$.

We are talking about the MLE, maximum likelihood estimator.

4 Thursday, 31 January 2013

First we minimized $\int l(\theta, s)n(\theta)d\theta$. The minimum is s^*_π. We figured that $s^*_\pi = E_\pi(\theta)$, the expectation of θ in the π distribution.

Example: Consider an experiment. Le $n = 5$ and have X_i is 0 or 1 for $i = 1, \cdots, 5$. So it is an indicator for a specific event (1 is have the allele, two is not).

Examples of Neymann factors. Ex 1. If n i.i.d. Bernoullis X_i, then the distribution of $(X_1, \cdots, X_n|\theta)$. Ex 2. You have $X_1, X_2, \cdots, X_n$ i.i.d. $N(\mu, \sigma_0^2)$ where variance is known and mean is unknown, then the likelihood is the product of the densities.

5 Thursday, 7 February 2013

Suppose for x, y for which $f(x|\theta)/f(y|\theta)$ is free of θ, we must have $T(x) = T(y)$. Suppose we have another sufficient statistic, suppose there exists T' and take x_1, x_2 in the same class with $T'(x_1) = T'(x_2)$. Because we are assuming T' is also sufficient, we have $f(x_1|\theta) = g'(T'(x_1), \theta)h'(x_1)$. And $f(x_2|\theta) = g'(T'(x_2), \theta)h'(x_2)$. Divide the two, and you have by the sufficiency class argument that the quotient is $h'(x_1)/h'(x_2)$. Then by the first sentence, $T(x_1) = T(x_2)$. We could say that T is a function of T'.

Example. Suppose X_i is iid distributed as $\text{Exp}(\lambda)$. Thus $\text{pr}(X_i|\lambda) = \lambda \exp(-\lambda X_i)$. Given a vector $(X_1, X_2, \cdots, X_n)$, its density is the product of the density for one random vector. Take a density of vector X^* and a density of vector Y^*, and you get

$$P(X^*|\lambda)/P(Y^*|\lambda) = \exp(-\lambda(\sum X_i - \sum Y_i))$$

If LHS is free of λ, then $\sum X_i = \sum Y_i$. The converse is true as well.

We have a case study. Expenses at time 1, then an indicator to see if we check them for E_2. The goal is to find $E(E_2)$. Our models are $P(E_1|\theta_1) = f(E_1, \theta_1), P(E_2|E_1, \theta_2) = g(E_2, E_1, \theta_2), P(I = 1|E_1, E_2) = \pi(E_1)$. The original expected value is $E(E_2) = \phi(\theta_1, \theta_2) = \int\int e_2 g(e_2, e_1, \theta_2) f(e_1, \theta) de_1 de_2$.

I dont quite understand the thing below. The likelihood in the above is

$$P(E_1, E_2^{ols}, I|\theta) = \prod_i \pi(E_i) \text{ blah blah}$$

Prior distribution. Bayes.

6 Tuesday, 12 February 2013

Suppose we want to relate Y to X. It could be as (1) $Y|_X \sim N(X\beta, \sigma^2)$. Or we could say (2) $E(Y|X, \beta) = X\beta$, but we leave all parts of $Y|_X$ unspecified.

Drawing of a lightning bolt blob of data.

Kostya is saying a lot of things. I do not understand what he is saying. He is not teaching in a context that I understand.

For just (2) without (1), no few functions are sufficient statistics. There won't be any reduction. It is harder to pass along to another scientist.

$(\hat{\beta}^{obs}|\beta, \theta - \beta$ which is n. Depends on the other unspecified parameters $) \sim N(\beta, \nu(\beta, n))$

from the slide

1. A few parameteres (estimands) can often describe the essence of a complex problem, and are more easily estimable.

2. Data in complex problems, however, usually cannot be reduced to a few statistics that are exactly sufficient for the estimands

3. Data can often be reduced to estimators that

   (a) have likelihood functions that are simpler to understand and communicate to others;

   (b) are approximately sufficient for the estimands

if the estimators that we keep (as reductions of the data) are close to sufficient for the estimand. Definition of sufficient = there is no parameter. Then the likelihoods from the display above will be more precise.

The estimand is what we try to estimate, it is β in this case.

There are two goals besides minimizing the loss function. (1) Try to reduce the data even if not by sufficient statistics so that you can understand the data better. (2) Be able to share the data with another scientist.

Precise would mean to look at the simpler likelihood, the RHS of the display should have a smaller variance.

Usually we can find the MLE by maximizing the $\log P(\mathbf{X}|\theta)$. Because then you get sums of logarithms.

Do it for binomial

Look at the MLE in some normal models. Let $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$. What is the MLE of μ or σ^2?

Recall

$$1/(2\pi\sigma^2)^{n/2} \exp\{-(\sum(X_i - \bar{X})^2 + n(\mu - \bar{X})^2)/2\sigma^2\}$$

Taking the log, you get

$$\log = -n/2 \log \sigma^2 - \sum(X_i - \bar{X})^2/2\sigma^2$$

Set it to 0. You get that $\sigma^2_{MLE} = \sum(X_i - \bar{X})^2/n$

7 Thursday, 14 February 2013

Note that from the multivariate normal, we get that $\hat{\beta}^{MLE} = \hat{\beta}^{OLS}$. This is because you have a quadratic form of a positive definite matrix in the likelihood function.

A numerical maximization is called Newton-Raphson. Also there is something similar called Fisher Scoring. Newton-Raphson depends on where you start. It can diverge if you are too far away. One algorithm that works well is EM (expectation maximization). It is a very smart algorithm because (generally you want to reduce the dimensions from many to few) this does the opposite. It expands the dimension of the problem to make it easier. Usually the dimension that is chosen is a natural dimension of something that you are missing.

Detour for PET, MRI.

THE MOST IMPORTANT CONCEPT IS LIKELIHOOD. THE SECOND MOST IMPORTANT IS MLE. The log odds of the probability is $\log(\frac{p}{1-p})$. EXPIT is a thing. It is defined as

$$\text{expit}(l) = \frac{e^l}{1+e^l}$$

Here, what about having $P(X_i|\theta_1)$ as $N(\mu_X, \sigma_X^2)$. Then $P(Y_i|X_i, \Theta_{Y|X}$ as $N(X_i\beta, \sigma_x^2)$. The distribution of X_i is known. And $\hat{E}^{MLE}(Y) = E(E(Y|X)) = \int yP(Y|X, \Theta_{Y|X})P(X|\Theta_n)dxdy$.

8 Tuesday, 19 February 2013

We are looking at sufficient statistics. We looked at MLE and at transformational invariance. Function of the MLE is the MLE of the functions. Looking at unbiased estimators will ultimately provide a continuity in sufficiency.

For any estimator that has an expectation equal to τ:

$$E[(\hat{\tau} - \tau)^2|\theta] = E[(\hat{\tau} - \tau^* + \tau^* - \tau)^2|\theta]$$

break it up and you get the variance and the bias^2.

You get an improvement in variance.

$$\text{var}(\hat{\tau}|\theta) = \text{var}\{E(\hat{\tau}|S,\theta)|\theta\} + E\{\hat{\tau}|S,\theta)|\theta\} \geq \text{var}(\tilde{\tau}|\theta)$$

the first term is Blackwellized estimator, $\tilde{\tau}$. This is theorem 7.3.17

Example. Say $X_1, \cdots, X_n$ are iid $\sim N(\mu, 1)$. Suppose we want to estimate $P(X_i < 0) = \Phi(-\mu)$ with an unbiased estimator. An obvious estimator for this is $1/n \sum X_i^* = \bar{X}^*$. Where $X_i^* = 1(X_i < 0)$. To improve $\bar{X}^*$, consider the sufficient statist of μ which is $\bar{X} = \sum X_i$.

Then $E(\bar{X}_i|\sum X_i)$ is unbaised MLE, and it will have smaller variance than $\bar{X}^*$. By symmetry,

$$E(\bar{X}^*|\sum X_i) = E(X_1^*|\sum X_i) \tag{1}$$

Notice that $X_1^* = 1(X_1 < 0)$ is just a count. We are trying to get a conditional distribution of the first observation given the sum. Notice, $(X_1, \sum X_i) \sim N((\mu, n\mu), [[1,1],[1,n]])$ So you can get the conditional distribution of $X_i | \sum X_i$. Eventually you get that the distro is $\Phi(-\bar{X}/\sqrt{1-1/n})$.

Theorem: Say for a fixed θ the loss $l(\theta, a)$ is convex. Then the estimator $\tilde{\theta}(S) = E(\hat{\theta}|S)$ has expected loss at most that of $\hat{\theta}$.

$$E(l(\theta, \hat{\theta}|S = s) \geq l(\theta, E(\hat{\theta}|S = s)) = l(\theta, \tilde{\theta}(s)).$$

9 Thursday, 21 February 2013

We have been talking about Blackwellization. We have been trying to improve the variance. We you take the expectation of an unbiased estimate conditioned on a sufficient statistic, you get better variance. It may sound counterintuitive that given fewer data you become more efficient. There was the theoretical result that you have to be dependent only on sufficient statistics (from the POV of decisions and data). Also, the improvement is similar to the principle of conditionality. For example, you have two scales, each with likelihoods. A has smaller variance, B has larger variance. You toss a coin to choose which scale you use to make a measurement. The conditionality principle says that your variance should only be defined on the scale that you actually use. Conditioning on a sufficient statistic can be seen as getting rid of noise.

Lehmann-Scheffe, if you have data in a model. There is a sufficient statistic that is "complete." When you take

$$\tau(\theta) = E\{g(S)|\theta\} \tag{2}$$

Suppose there exists another estimator, $Q(X) : E(Q(X)|\theta) = \tau(\theta)$. And $\text{var}(Q(X)|\theta) < \text{var}(g(S)|\theta)$ for some θ. Then the Blackwellization $\tilde{g(S)} = E(Q(X)|S)$ is better than $Q(X)$ in the sense that

$$\text{var}(\tilde{g(\theta}(S)|\theta) \leq \text{var}(Q(X)|\theta) < \text{var}(g(S)|\theta)$$

Then we have $\tilde{g(\theta}(S), g(S)$ unbiasted for $\tau(\theta)$ so

$$E\{\tilde{g(\theta}(S) - g(S)|\theta\} = 0$$

so if S is complete, then the two functions in the expectation above have to be the same almost surely. That contradicts the fact that the one has a variance strictly less than the other.

EXAMPLE if X is a binomial(n, θ) and $\theta \in (0, 1)$. Then $P(X|\theta) = (n, x)\theta^x(1-\theta)^{n-x} = (n, x)\exp\{x\log(\theta) + (n-x)\log(1-\theta)\}$. So going across, we see the $t_1(x), \gamma_1(\theta), t_2(x), \gamma_2(\theta)$. And it is also equal to

$$(n, x)(1-\theta)^n \exp\{x\log(\frac{\theta}{1-\theta})\}$$

Notice that this is two different ways to represent the same thing as a member of the exponential family. So in general we are considering the representation where the functions are linearly independent. We will represent the ts and the γs such that if

$$\sum_{j=1}^{k} \gamma_j(\theta)\alpha_j = \text{ constant in } \theta, \text{ then } \alpha_j = 0, \text{ for all } j$$

and the same thing for the ts.

EXAMPLE: $X \sim \beta(\alpha, \beta)$ with $\alpha > 0$ and $\beta > 0$. then if we have $x \in (0, 1)$

$$P(X|\theta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\beta(\alpha, \beta)}$$

is it an exponential family? YES!

$$P(X|\theta) = I_{x\in(0,1)} \frac{\exp((\alpha-1)\log(x) + (\beta-1)\log(1-x)}{\beta(\alpha, \beta)}$$

A counter example is the uniform distribution.

It is easy to spot the minimal statistic, and it is easy to see if it is complete. For a general exponential family, then the sufficient statistic is just the sum of the $t_i(x_j)$ functions, sum over j.

One way to show what is a minimal statistic. Take the vector of observations, find the equivalence class of two observations whose ratio does not depend on θ. Take two vectors x, y, and look at the likelihood ratio,

$$P(X|\theta)/P(Y|\theta) = \frac{\prod h(X_i) \exp[\sum \sum \gamma_j(\theta) t_j(X_i)]}{\prod h(Y_i) \exp[\cdots t_j(Y_i)]}$$

And we need that the following does not vary with θ

$$\exp[\sum \gamma_j(\theta)\{\sum t_j(X_i) - \sum t_j(Y_i)\}]$$

So we must have

$$\sum_i t_j(X_i) = \sum_i t_j(Y_i)$$

for all j

Tuesday, 26 February 2013

Hey there! We are back. For statistic T, if

$$T(x) = T(y) = \frac{f(x|\theta)}{f(y|\theta)} \qquad \text{free of } \theta$$

Then T is a minimally sufficient statistic. The converse is also true. And therefore $f(x| \cdot)$ is minimally sufficient.

Look at the normal distribution as a member of the exponential family. The set of (μ, σ^2) contains an open set. And that leads to some results that we like.

Thursday, 7 March 2013

The Cramer-Rao bound for unbiased estimators is

$$\left(\frac{d\tau}{d\theta}(\theta_1,\theta_2)\right)' I^{-1}(\theta_1,\theta_2)\left(\frac{d\tau}{d\theta}(\theta_1,\theta_2)\right) \tag{3}$$

In the derivation we needed that $E(S_j|\theta)=0$ and that $E(WS_j|\theta)=\frac{dE(W|\theta)}{d\theta_j}$. The idea is that you can interchange the derivative and integral.

We have defined the information matrix $I(\theta)=\text{var}(S_1,S_2|\theta)=$

$$\begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \tag{4}$$

This is called Fisher's information matrix for θ_1,θ_2.

EXAMPLE Suppose we want to estimate θ_1. So we have that $\tau(\theta_1,\theta_2)=\theta_1$ and $d\tau/d\theta(\theta_1,\theta_2)=(1,0)$. If we knew θ_2, the CR bound would be $(I_{11})^{-1}$. This (the Cramer-Rao bound) is the variance of any unbiased projection W on S_1. Suppose we do not know θ_2 then the CR bound would be given by the first term in (3). This would be the $1,1$ element of the $I(\theta)^{-1}$ matrix.

It is useful to see what the $1,1$ element is of the inverse $[I^{-1}]_{11}$.

$$[I^{-1}(\theta)]_{11}=(I_{11}-I_{12}I_{22}^{-1}I_{21})^{-1} \tag{5}$$

$$=\text{var}(S_1-I_{12}I_{22}^{-1}S_2)^{-1} \tag{6}$$

Consider two skew vectors, S_1 and S_2, then $I_{12}I_{22}^{-1}S_2$ is the projection of S_1 onto S_2. Notice $S_1^{eff}=S_1-I_{12}I_{22}^{-1}S_2$. This is the efficient score. The information decreases, so the bound on the variance increases. The bound is from Blackwellizing on the scores.

$$\log \text{pr}(X|\theta_1,\theta_2)=\log f_1(X|\theta_1)+\log f_2(X|\theta_2) \tag{7}$$

So in the cross terms, I_{12} we get

$$\frac{d\log(X|\theta_1,\theta_2)}{d\theta_1 d\theta_2} \tag{8}$$

EXAMPLE

We have X then an indicator I whether or not to sample, then the ones that were sampled, we have Y, the value that was observed (this is the three bar thing). For estimating, we have

$$E(Y)=E(Y|\theta_{Y|X},\theta_X=\int Y\,\text{pr}(Y|X,\theta_{Y|X})\text{pr}(X|\theta_X)dx$$

Remember we had the likelihood under an ignorable design. That was

$$\prod_i P(X|\theta_x)\prod_{i:I_i=1}\text{pr}(Y_i|X_i,\theta_{Y|X}) \tag{9}$$

If we obtain $\hat{\theta_X},\hat{\theta_{Y|X}}$ as MLEs then their variances will be $\text{var}(S_{\theta_X},S_{\theta_{Y_X}}|\theta)^{-1}$ Now lets say we want to estimate $\tau(\theta_X,\theta_{Y|X})$ by $\tau(\hat{\theta_X},\hat{\theta_{Y|X}})$

EXAMPLE

You want to estimate the survival time for a cohort. Surgery time to death, or start of chemotherapy to death. You get data, T_i times of living. The goal is to find $P(T_i > t)$ where t could be something like 1 year. Note that $P(T_i > t)$ can be estimated using the Kaplan Meier.

10 Tuesday 12 March 2013

We have X, I, Y. We have X is the starting population. I is the indicator that we observe a particular x_i. Y is the values for the observations. The design is

1. Here (X_i, Y_i, I_i) are i.i.d.
2. Here $\text{pr}(I_i = 1|X_i, Y_i) = \text{pr}(I_i = 1|X_i) = \pi(X_i)$

The model is $\text{pr}(X_i|\theta_X) = f(x, \theta_X)$ and $\text{pr}(Y_i|X_i, \theta_{Y|X}) = g(Y, X, \theta_{Y|X})$. The likelihood is

$$\prod_i \text{pr}(X_i|\theta_X) \prod_{I=i} \text{pr}(Y_i|X_i, \theta_{Y|X} \prod \pi(X_i)^{I_i}(1 - \pi(X_i))^{1-I_i} \tag{10}$$

For our notation we have $\text{pr}(X_i = x_i, I_i = 1, Y_i = y_i|\theta) = \text{pr}(X_i = x_i)\text{pr}(Y_i = y_i|X_i = x_i)\text{pr}(I_i = 1|Y_i = y_i, X_i = x_i)$ by conditional probability. By the design of the matrix, the third term is $\pi(x_i)$.

Now let's say we have $I_i = 0$. Then $\text{pr}(X_i = x_i, I_i = 0|\theta) = \text{pr}(X_i = x_i)\text{pr}(I_i = 0|X_i = x_i)$. The second term contributes to the third term of (10). More generally, you could have written $\int \text{pr}(X_i = x_2, I_i = 0, Y - i = y|\theta)dy$ by the law of total probability. It is $\int \text{pr}(X_i = x_2)\text{pr}(Y_i = y|X_i = x_2)\text{pr}(I_i = 0|X_i = x_2, Y_i = y)dy$.

In homework 2, the last problem asks about an inference given a correct and an incorrect model. Check what happens when the sample size increases, and you will see that the likelihood changes to match the parameter (and the incorrect model gives you something different).

Use the theory of exponential families to get the MVUE, and then see if the family is complete.

We have

$$I = \text{var}\begin{bmatrix} S_\beta \\ S_(\sigma^2) \end{bmatrix} = \tag{11}$$

3x3 , 3x1 // 1x3, 1x1 matrix.

If we have a $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$ i.i.d. What is a CRb for μ? It is $\text{var}(S_\mu)$ because here we don't have covariances in the off diagonal to worry about. It is equal to

$$\text{var}(-\frac{\sum(Y_i - \mu)}{\sigma^2}) = \frac{n\sigma^2}{\sigma^4} = \frac{n}{\sigma^2} \tag{12}$$

so $\text{Info}(\mu) = 1/\text{var}(S_\mu) = \sigma^2/n$. Some limitations of the unbiased estimators, it is the case that they do not always exist. And, they are most often no admissable.

The most famous example of all of this is the James and Stein's Estimator (you can read his, Stein, paper). It is really an ANOVA.

Consider treating cells with one of K treatments. We are interested in estimating the average time, θ_K, for cells to replicate under each treatment, k. We summarize the data by the averages X_k in each treatment and assume that (1) CLT (2) we wish to evaluate any estimator $\theta = \{\theta_k, k = 1, \cdots, K)\}$ by the combined squared error loss

$$L(\hat{\theta}, \theta) = E(\sum_K (\hat{\theta_k} - \theta_k)^2 | \theta) \tag{13}$$

Let's say we have three harsh treatments for mice. The true means for the outcomes are $\theta_1, \theta_2, \theta_3$. You have a normal distribution centered at each of the θs (think complicated bar chart here). ANOVA assumes normality and constant variance. Assume that here.

Talking about the Stein estimator. It is a convex combination (think BAYES) of the average and a more parimonious model that says everything is the same.