University of Illinois
Spring 2017

Econ 574

Department of Economics
Roger Koenker

Lecture 8
"Happy Families Are All The Same"

In this lecture we consider the exponential family of densities which provide an idealized framework for MLE.

Def: A family $\mathcal{F}_\theta$ of parametric models is a *one-parameter exponential family*, $(1pxf)$, if there exist functions $c(\theta), d(\theta)$ on Θ and real-valued functions T, S on $\Re$, and a set A such that elements of $\mathcal{F}_\theta$ have densities

$$f(z,\theta) = \exp\{c(\theta)T(z) + d(\theta) + S(z)\}I_A(z)$$

Recall: $I_A(z) = \begin{cases} 1 & z \in A \\ 0 & z \notin A \end{cases}$ so this term merely constrains the support of f, but note the support, A, is explicitly independent of θ. E.g., $A = \{0, 1, 2, \dots\}$ in some cases, but the support can't depend upon θ!!

Random Sampling from $1pxf$'s

Consider $Z_1, \dots, Z_n$ iid from a $1pxf$ then the likelihood may be written as

$$\mathcal{L}(\theta|z) = \prod_{i=1}^n f(z_i|\theta) = \exp\{c(\theta)\sum_{i=1}^n T(z_i) + nd(\theta) + \sum_{i=1}^n S(z_i)\}\prod_{i=1}^n I_A(z_i)$$

so joint distribution is also a $1pxf$ and $T = \sum_{i=1}^n T(z_i)$ is sufficient for θ.

Often it is useful to treat $c(\theta)$ as "the parameter of interest", so we call $\eta = c(\theta)$ "the natural parameter" and write,

$$f(z|\eta) = \exp\{\eta T(z) + d_0(\eta) + S(z)\} \qquad \text{where } d_0(\eta) = d(c^{-1}(\eta))$$

Moments of $T(z)$ in $1pxf$'s

In order to evaluate the behavior of $T(z)$ we would like to be able to evaluate $ET(Z)$ and $V(T(Z))$. This can be done easily with the following trick. Consider the identity

$$\int f(z|\eta)dz = 1$$

since the support is independent of η we can differentiate under the integral to get,

$$\begin{aligned} \frac{d}{d\eta}\int f(z|\eta)dz = 0 \quad &\Rightarrow \quad \int (T(z) + d_0'(\eta)f(z|\eta)dz = 0 \\ &\Rightarrow \quad E_\eta T(Z) = -d_0'(\eta) \end{aligned}$$

differentiating again yields,

$$\begin{aligned}
\frac{d}{d\eta}\int (T(z)+d_0'(\eta))f dz = 0 & \\
&\Rightarrow \int (d_0''(\eta))f dz + (T+d_0')^2 f dz = 0\\
&\Rightarrow E(T(Z)+d_0'(\eta)) = -d_0''(\eta)\\
&\Rightarrow E(T(Z)-ET(Z))^2 = V(T(Z)) = -d_0''(\eta)
\end{aligned}$$

This can be done with mgf's, but is much less pleasant!

Example

(1) Poisson

$$f(x,\theta) = \frac{\theta^x e^{-\theta}}{x!} I_A(x) \qquad\qquad A = \{0,1,\dots\} \text{ so } f(x,\theta) = \exp\{x\log\theta - \theta - \log(x!)\} I_A(x)$$

So the MLE for θ, $\hat{\theta} = \bar{x}$, is easily obtained,

$$l(\theta) = \log(\theta)\sum x_i - n\theta - \sum \log x_i!$$

$$\nabla l(\theta) = \theta^{-1}\sum x_i - n = 0$$

What is “happiness”? Digression on the Cramér-Rao Inequality

> The Cramér-Rao inequality is ”much less deep than a random line from Ramanujan's notebooks”
> C.R. Rao, Hip Pocket Restaurant, Champaign, December, 1983.

Thm: Let $T(z)$ be any estimator of θ for a model with likelihood $f(z|\theta)$. Assume

(i): The support, $\mathcal{Z} = \{z | f(z|\theta) > 0\}$, doesn't depend on θ

(ii): For all $z \in \mathcal{Z}$ and $\theta \in \Theta \quad |\partial \log f/\partial\theta| < \infty$

(iii): For *any* $T(z)$ such that $E|T(z)| < \infty$

$$\frac{\partial}{\partial\theta}\int_{\mathcal{Z}} T(z) f(z|\theta) dz = \int T(z)\frac{\partial f(z|\theta)}{\partial\theta} dz$$

Let $E_\theta T(z) = t(\theta)$ and $V_\theta(\partial \log f/\partial\theta) = I(\theta)$ then $V_\theta T(z) \geq [t'(\theta)]^2/I(\theta)$

Pf: All integrals over the full support $\mathcal{Z}$;

(1): $\int f dz = 1$

(2): $\int T f dz = t(\theta)$

using (iii), setting $T(z) \equiv 1$ in $1'$, with $l \equiv l(z|\theta) \equiv log f(z|\theta)$,

($1'$): $\frac{\partial}{\partial\theta}\int f dz = \int \frac{\partial}{\partial\theta} f dz = \int \frac{\partial}{\partial\theta} l \cdot f dz = 0$

($2'$): $\frac{\partial}{\partial\theta}\int T \cdot f dz = \int T \cdot \frac{\partial}{\partial\theta} f dz = \int T \frac{\partial}{\partial\theta} l \cdot f dz = t'(\theta)$

Now set $X_1 = T(Z)$ and $X_2 = \frac{\partial}{\partial\theta}\log f(Z|\theta)$ and rewrite again as,

($1''$): $EX_2 = 0$

($2''$): $EX_1X_2 = t'(\theta)$

So Cov $(X_1, X_2) = EX_1X_2 - EX_1EX_2 = t'(\theta)$ and since

$$|\rho(X_1, X_2)| = \frac{|\text{Cov }(X_1, X_2)|}{\sqrt{V(X_1)V(X_2)}} \leq 1$$

we have,

$$|t'(\theta)| \leq \sqrt{V(T(Z))V(\frac{\partial}{\partial\theta}\log f)}$$

or

$$V(T(Z)) \geq \frac{(t'(\theta))^2}{I(\theta).} \qquad \Box$$

Cor: Suppose $ET(Z) = \theta$, then $V(T(Z)) \geq I^{-1}(\theta)$

Pf: $ET(z) = t(\theta) = \theta \Rightarrow t'(\theta) = 1.$

Information

Cor: Suppose $Z_1, \ldots, Z_n$ is a random sample from $f(z|\theta)$, then

$$V(T(Z)) \geq \frac{(t'(\theta))^2}{nI_1(\theta)}$$

Pf:

$$I(\theta) = V(\frac{\partial}{\partial\theta}\log f(Z|\theta)) = V(\sum\frac{\partial}{\partial\theta}\log f(Z_i, \theta)) = nI_1(\theta).$$

Digression à la Pitman (1979) §3.1 on Fisher Information.

The Hellinger distance between f, f_0 is

$$\rho^2(f, f_0) = \int(\sqrt{f} - \sqrt{f_0})^2 d\mu = 2 - 2\int\sqrt{ff_0}d\mu.$$

Clearly $\rho(f, f_0) = 0$ if $f = f_0$ and $\rho^2(f, f_0) = 2$ if the supports of f and f_0 don't intersect.

Consider

$$\frac{\rho^2}{(\theta - \theta_0)^2} = \int\frac{(\sqrt{f} - \sqrt{f_0})^2}{(\theta - \theta_0)^2}d\mu$$

then, if $\sqrt{f}$ has a $(a.e.\mu)$ θ-derivative at θ_0, then since

$$\left(\frac{d\sqrt{f}}{d\theta}\right)^2_{\theta=\theta_0} = \left(\frac{1}{2}\frac{f'}{\sqrt{f}}\right)^2_{\theta=\theta_0} = \frac{(f_0')^2}{4f_0}$$

We have

$$\lim_{\theta\to\theta_0}\frac{\rho^2}{(\theta - \theta_0)^2} = \int\frac{(f_0')^2}{4f_0^2}f_0 d\mu = \frac{1}{4}I(\theta_0)$$

or

$$\lim_{\theta\to\theta_0}\frac{\rho}{|\theta - \theta_0|} = \frac{1}{2}\sqrt{I(\theta_0)}$$

Pitman calls the lhs the sensitivity of the family $\mathcal{F}_\theta$ to perturbations of θ at θ_0. Clearly Fisher information measures the rate at which f diverges from f_0 in Hellinger distance as θ diverges from θ_0.

Thm: (BD 4.3.2) If $\mathcal{F}_\theta$ satisfies the conditions of the CRLB, and there exists T^* such that $ET^* = t(\theta)$ and $VT^* = (t'(\theta))^2/I(\theta)$, i.e., T^**achieves* CRLB, for all $\theta \in \Theta$, then $\mathcal{F}_\theta$ is a $1pxf$.

Remark: First, we note that the sufficient statistic $T(z)$ achieves the CRLB. We have already seen that

$$\begin{aligned} ET &= -d_0'(\eta) \\ VT &= -d_0''(\eta) \end{aligned}$$

CRLB implies that for any statistic $R(z)$ with $ER = r(\eta)$

$$VR \geq \frac{(r'(\eta))^2}{T(\eta)}$$

so for T,

$$\begin{aligned} V(T) &\geq (-d_0''(\eta))^2/(-d_0''(\eta)) \\ &= -d_0''(\eta) \end{aligned}$$

which is achieved.

Pf: Equality in the CRLB implies from the correlation inequality that

$$\frac{\partial}{\partial\theta} \log f(z|\theta) = a(\theta)T^* + b(\theta)$$

for all $\theta \in \Theta$. Integrating and then exponentiating gives a density in $1pxf$ form.

Compound Decisions, Bayes Rules and Exponential Families

Suppose that we observe the following process:

$$Y_i \sim \mathcal{N}(\mu_i, \sigma_0^2) \quad i = 1, \cdots, n,$$

and would like to estimate *all* the μ_i's, subject to squared error loss. We have already seen that the James Stein estimator,

$$\hat{\mu}_i = (1 - \frac{n-2}{S})Y_i$$

is one way to do this by shrinking all the unbiased estimates, Y_i, toward zero. We may take the following Bayesian perspective: if we thought that the μ_i's were drawn iid-ly from a distribution F, then the observed Y_i's would have the convolution density,

$$g(y) = \int \varphi_0(y - \mu)dF(\mu),$$

where we let $\varphi_0(u) = \phi(u/\sigma_0))/\sigma_0$. If we knew F, so we also knew g, then the Bayes rule for estimating the μ's would be:

$$\delta(y) = y + g'(y)/g(y).$$

This result is called Tweedie's formula by Efron (2011). We will show that it follows very simply from an exponential family argument, but before doing that I'd like to make an argument for its plausibility by showing its connection to the James-Stein estimator.

Suppose that we thought that the mixing distribution, F, was $\mathcal{N}(0, \sigma_1^2)$ then clearly the mixture distribution is,

$$g(y) = \phi(y/\sqrt{\sigma_0^2 + \sigma_1^2})/\sqrt{\sigma_0^2 + \sigma_1^2}$$

and therefore,

$$\frac{g'(y)}{g(y)} = \frac{-y}{\sigma_0^2 + \sigma_1^2}.$$

Thus, we obtain from Tweedie's formula the decision rule,

$$\delta(y) = (1 - \frac{1}{\sigma_0^2 + \sigma_1^2})y.$$

If we now admit that we don't really know σ_1^2 and just decide to use the natural estimator for the variance of the Y_i's based on,

$$S = \sum Y_i^2 \sim (\sigma_0^2 + \sigma_1^2)\chi_n^2,$$

noting tht if we replace $\sigma_0^2 + \sigma_1^2$ by S, and recall (!) that the expectation of the reciprocal of a χ_n^2 is $1/(n-2)$, we obtain precisely the James Stein estimator.

Now suppose that we no longer want to assume normality for F, where does the Tweedie formula really come from? Recall that for squared error loss the Bayes rule requires us to choose the expectation of μ based on the posterior. To compute this expectation for our normal context it is actually easier to consider the more general setting in which Y comes from the exponential family model,

$$\varphi(y, \theta) = m(y)e^{y\theta}h(\theta).$$

Note that in this general framework we no longer even have the location shift form, so the mixtures are no longer necessarily convolutions, but not to worry, we will soon revert back to special case of the normal model. Let's denote the random θ as Θ, so we would like to compute:

$$\begin{aligned}\delta(y) &= \mathbb{E}(\Theta|y)\\ &= \int \theta\varphi(y,\theta)dF(\theta)/\int \varphi(y,\theta)dF(\theta)\\ &= \int \theta e^{y\theta}h(\theta)dF(\theta)/\int e^{y\theta}h(\theta)dF(\theta)\\ &= \frac{d}{dy}\log\int e^{y\theta}h(\theta)dF(\theta)\\ &= \frac{d}{dy}\log(g(y)/m(y))\end{aligned}$$

Now returning to the normal case,

$$\varphi(y,\theta) = \phi(y-\theta) = K\exp(-y^2/2)\exp(y\theta)\exp(-\theta^2/2),$$

so $m(y) = \exp(-y^2/2)$ and our log derivative yields the Tweedie formula.

Invariance of CRLB

It is important to remember that $I(\theta)$ depends upon the particular parameterization employed. Thus, in the model $f(z|\theta)$, suppose that $\theta = h(\eta)$ we can express the information about η as

$$I^*(\eta) = I(h(\eta))(h'(\eta))^2$$

where $I(\cdot)$ denotes the Fisher information with respect to to θ and I^* with respect to to η. To see this note

$$\begin{aligned}I^*(\eta) &= V(\frac{\partial}{\partial\eta}f(z|h(\eta)))\\ &= V(\frac{\partial}{\partial\theta}f(z|\theta)\cdot h'(\eta))\\ &= I(\theta)\cdot(h'(\eta))^2\end{aligned}$$

But note also that if $\theta = h(\eta)$ then since

$$\frac{\partial}{\partial \eta} r^*(h(\eta)) = r'(\theta)h'(\eta)$$

we have

$$\frac{(r'(\theta))^2}{I(\theta)} = \frac{(r^{*\prime}(\eta))^2}{I^*(\eta)} \qquad \Box$$

Multiparameter Extensions

For problems with $\Theta \subset \Re^p$ we have

$$\begin{aligned} I(\theta) &= (I_{ij}(\theta)) \\ I_{ij}(\theta) &= E\left(\frac{\partial}{\partial\theta_i} l(z|\theta) \cdot \frac{\partial}{\partial\theta_j} l(z|(\theta)\right) \end{aligned}$$

Again differentiating under the integral gives

$$E\frac{\partial}{\partial\theta}\log f = 0$$

so

$$I_{ij} = \text{Cov }\left(\frac{\partial}{\partial\theta_i} l, \frac{\partial}{\partial\theta_j} l\right) = -E\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j} l\right)$$

This identity is very important.

p-parameter exponential family

$$f(z|\eta) = \exp\{\sum_{i=1}^p \eta_i T_i(z) + d_0(\eta) + S(z)\} I_A(z)$$

Obviously, $T(z) = (T_i(z))$ is jointly sufficient for η.

It is straightforward to show as in the $1pxf$ case that

$$\begin{aligned} E(T_i(z)) &= -d_0^{(i)}(\eta) \\ \text{Cov }(T_i, T_j) &= -d_0^{(ij)}(\eta) \end{aligned}$$

where $d_0^{(i)} = \frac{\partial d_0(\eta)}{\partial \eta_i}$ and $d_0^{(ij)} = \frac{\partial^2 d_0(\eta)}{\partial\eta_i\partial\eta_j}$.

But now that is no presumption that the CRLB is attained. In p dimensions we need to adapt the result somewhat and we simply state it without proof. See, e.g., Lehmann for details.

Thm: Suppose $\mathcal{F}_\theta$ is a family indexed by $\theta \in \Theta \subset \Re^p$. Assume

(i): the set $\mathcal{Z} = \{z | f(z|\theta) > 0\}$ doesn't depend on θ.

(ii): For all $z \in \mathcal{Z}$ and $\theta \in \Theta$, $\partial \log f / \partial\theta_i < \infty \quad i = 1, \ldots, p$.

(iii): For any scalar $R(z)$ such that $E|R| < \infty$ one can differentiate under the integral sign.

Then, setting $\alpha = \nabla E_\theta(R(Z))$, we have, $V(R(Z)) \geq \alpha' I^{-1}\alpha$.

References

Pitman, E.J.G. (1979) *Some Basic Theory for Statistical Inference*, Halstad Press.