STAT 528 - Advanced Regression Analysis II

Exponential family theory

Daniel J. Eck
Department of Statistics
University of Illinois

Agenda for today

- Course software and GitHub
- Go over basics of exponential family theory
- Cover regular full exponential families and their properties
- Discuss identifiability

Exponential family

An *exponential family of distributions* is a parametric statistical model having log likelihood that takes the form

$$l(\theta) = \langle y, \theta \rangle - c(\theta), \tag{1}$$

where y is a vector statistic and θ is a vector parameter, and

- $\langle y, \theta \rangle$ is the usual inner product,
- $c(\theta)$ is the cumulant function.

This uses the convention that terms that do not contain the parameter vector can be dropped from a log likelihood; otherwise additional terms also appear in (1).

When the log likelihood can be expressed as (1) we say that y is the **canonical statistic** and θ is the **canonical parameter**.

Example: Binomial distribution

Let $X \sim \text{Binomial}(n,p)$ where $0 < p < 1$. We can write the log probability mass function for X

$$\begin{aligned} l(p) &= \log\left(\binom{n}{x}\right) + x\log(p) + (n-x)\log(1-p) \\ &\propto x\log(p) + (n-x)\log(1-p) \end{aligned}$$

in exponential family form

$$l(\theta) = \langle y, \theta \rangle - c(\theta).$$

Densities

Let w represent the full data, then the densities have the form

$$f_\theta(w) = h(w) \exp\left(\langle Y(w), \theta \rangle - c(\theta)\right) \tag{2}$$

and the word “density' ' here can refer to a PMF, PDF, or to a density with respect to a positive measure.

The $h(w)$ arises from any term not containing the parameter that is dropped in going from log densities to log likelihood as we saw on the previous slide.

The function h has to be nonnegative, and any point w such that $h(w) = 0$ is not in the support of any distribution in the family.

Example: Binomial distribution

Let $X \sim \text{Binomial}(n,p)$ where $0 < p < 1$. We can write the probability mass function for X

$$f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

as an exponential family density

$$f_\theta(w) = h(w) \exp\left(\langle Y(w), \theta \rangle - c(\theta)\right).$$

Example: Normal distribution

Let $W \sim N(\mu, \sigma^2)$. Then we can write

$$f_{\mu,\sigma^2}(w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(w-\mu)^2}{2\sigma^2}\right)$$

as an exponential family density

$$f_\theta(w) = h(w) \exp\left(\langle Y(w), \theta\rangle - c(\theta)\right),$$

where

$$c(\theta) = \frac{1}{2}\left(\frac{\theta_1^2}{2\theta_2} - \log(2\theta_2)\right).$$

Cumulant functions

Being a density, (2) must sum, integrate, or sum-integrate to one. Hence,

$$\begin{aligned} 1 &= \int f_\theta(w)dw \\ &= \int \exp\left(\langle Y(w), \theta\rangle - c(\theta)\right) h(w)dw \\ &= \exp\left(-c(\theta)\right) \int \exp\left(\langle Y(w), \theta\rangle\right) h(w)dw. \end{aligned}$$

Rearranging the above implies that

$$c(\theta) = \log\left(\int \exp\left(\langle Y(w), \theta\rangle\right) h(w)dw\right).$$

The cumulant function is the log Laplace transformation corresponding to the *generating measure* given by

$$\lambda(dw) = h(w)dw$$

when the random variable is continuous. Under this formulation

$$c(\theta) = \log\left(\int \exp\left(\langle Y(w), \theta\rangle\right) \lambda(dw)\right).$$

In our log likelihood based definition of the exponential family (1), the dropped terms which do not appear in the log likelihood are incorporated into the counting measure (discrete distributions) or Lebesgue measure (continuous distributions).

Full families

Define

$$\Theta = \{\theta : c(\theta) < \infty\}. \tag{3}$$

Then (1) and (2) define a distribution for all $\theta \in \Theta$.

We say an exponential family is *full* if its canonical parameter space is (3). Many commonly used statistical models are full exponential families.

Moment generating functions

We no longer fuss about $Y(w)$ and will suppress w when writing Y. We still mention the function h in (2) which is now derived with respect to Y instead of w.

The moment generating function of the canonical statistic, if it exists, is given by

$$\begin{aligned} M_\theta(t) &= \mathrm{E}_\theta\left(e^{\langle Y, t\rangle}\right) \\ &= e^{c(\theta+t)-c(\theta)}. \end{aligned} \tag{4}$$

The moment generating function exists if θ is an interior point of the full canonical parameter space (3).

By the theory of moment generating functions, if the moment generating function exists, then moments of all orders exist and ordinary moments are given by the derivatives of $M_\theta(t)$ evaluated at zero.

In particular,

$$\begin{aligned} \mathsf{E}_\theta(Y) &= \nabla M_\theta(0) = \nabla c(\theta) \\ \mathsf{E}_\theta(YY^T) &= \nabla^2 M_\theta(0) = \nabla^2 c(\theta) + [\nabla c(\theta)][\nabla c(\theta)]^T. \end{aligned}$$

Cumulant generating function

A log moment generating function is called a *cumulant generating function* and its derivatives evaluated at zero are called the *cumulants* of the distribution.

For θ in the interior of the full canonical parameter space Θ, the cumulant generating function corresponding to the canonical statistic is

$$k_\theta(t) = c(t + \theta) - c(\theta), \tag{5}$$

where $c(\theta)$ is the cumulant function corresponding to the exponential family in canonical form.

The derivatives of $k_\theta(t)$ evaluated at 0 are the same as the cumulant function c evaluated at θ. The first and second cumulants of the canonical statistic are

$$\begin{aligned} \nabla c(\theta) &= \mathsf{E}_\theta(Y) \\ \nabla^2 c(\theta) &= \mathsf{E}_\theta(YY^T) - [\mathsf{E}_\theta(Y)]\,[\mathsf{E}_\theta(Y)]^T = \mathsf{Var}_\theta(Y). \end{aligned} \tag{6}$$

Regular exponential families

An exponential family is regular if its full canonical parameter space (3) is an open set so that the moment and cumulant generating functions exist for all θ.

The formulas in the preceding section hold for all θ.

Nearly every exponential family that arises in applications is regular. We will not discuss non-regular exponential families.

Example: Binomial distribution

The Binomial distribution with the standard parameter space $0 < p < 1$ written in canonical form is a regular full exponential family.

We already saw that the loglikelihood for the Binomial distribution can be written as

$$l(\theta) = \langle Y, \theta \rangle - c(\theta).$$

The canonical parameter θ for this exponential family is $\theta = \log\left(\frac{p}{1-p}\right)$ and this implies that $c(\theta) < \infty$ for all $\theta \in \mathbb{R}$ and that Θ is open.

Take home message

Let Y be a regular full exponential family in canonical form. Then Y has log likelihood given by

$$l(\theta) = \langle Y, \theta \rangle - c(\theta),$$

and

$$\begin{aligned} \mathrm{E}_\theta(Y) &= \nabla c(\theta), \\ \mathrm{Var}_\theta(Y) &= \nabla^2 c(\theta). \end{aligned}$$

Identifiability

A statistical model is *identifiable* if any two distinct parameter values correspond to distinct distributions.

An exponential family fails to be identifiable if there are two distinct canonical parameter values θ and ψ such that the density (2) of one with respect to the other is equal to one with probability one.

This happens if $Y^T(\theta - \psi)$ is equal to a constant with probability one.

And this says that the canonical statistic Y is concentrated on a hyperplane and the vector $\theta - \psi$ is perpendicular to this hyperplane.

Conversely, if the canonical statistic Y is concentrated on a hyperplane

$$H = \{y : y^T v = a\} \tag{7}$$

for some non-zero vector v, then for any scalar s

$$c(\theta + sv) = sa + c(\theta),$$

which immediately implies that

$$l(\theta + sv) = l(\theta).$$

Theorem

An exponential family fails to be identifiable if and only if the canonical statistic is concentrated on a hyperplane. If that hyperplane is given by (7) *and the family is full, then* θ *and* $\theta + sv$ *are in the full canonical parameter space and correspond to the same distribution for every canonical parameter value* θ *and every scalar* s.

- The direction sv along a vector v in the parameter space such that θ and $\theta + sv$ always correspond to the same distribution is called a *direction of constancy*.
- The theorem says that v is such a vector if and only if $Y^T v$ is constant with probability one.
- It is clear from this that the set of all such vectors is closed under vector addition and scalar multiplication, hence is a vector subspace. This subspace is called the *constancy space* of the family.

Note

It is always possible to choose the canonical statistic and parameter so the family is identifiable.

Y being concentrated on a hyperplane means some components are affine functions of other components with probability one, and this relation can be used to eliminate components of the canonical statistic vector until one gets to an identifiable choice of canonical statistic and parameter.

But this is not always advisable. Prematurely enforcing identifiability may complicate many theoretical issues.

Example: Multinomial distribution

We will show that a nonidentifiable parameterization allows for a relatively routine argument to show that the multinomial distribution can be written in canonical form.

First, some background on ratios of densities:

When we look at a ratio of two exponential family densities with canonical parameter vectors θ and ψ, the $h(w)$ term cancels, and

$$f_{\theta;\psi}(w) = \frac{f_\theta(w)}{f_\psi(w)} = e^{\langle y, \theta - \psi \rangle - c(\theta) + c(\psi)} \tag{8}$$

is a density of the distribution with canonical parameter θ taken with respect to the distribution with canonical parameter ψ (a Radon-Nikodym derivate in probability theory).

In the same vein as (8), we obtain the identity

$$c(\theta) = c(\psi) + \log\left(\mathrm{E}_\psi\left(e^{\langle Y,\theta-\psi\rangle}\right)\right) \tag{9}$$

Then (9) gives the following for the multinomial distribution

$$\begin{aligned} c(\theta) &= c(\psi) + \log\left(\mathrm{E}_\psi\left(e^{\langle Y,\theta-\psi\rangle}\right)\right) \\ &= c(\psi) + n\log\left(\sum_{i=1}^{d} p_i e^{\theta_i-\psi_i}\right), \end{aligned}$$

where the last equality follows from the multinomial theorem.

Then (8) gives

$$\begin{aligned} f_\theta(y) &= f_\psi(y) e^{\langle y, \theta - \psi \rangle - c(\theta) + c(\psi)} \\ &= \binom{n}{y} \left(\prod_{i=1}^{d} \left[p_i e^{\theta_i - \psi_i} \right]^{y_i} \right) \left(\sum_{i=1}^{d} p_i e^{\theta_i - \psi_i} \right)^{-n} \\ &= \binom{n}{y} \prod_{i=1}^{d} \left(\frac{p_i e^{\theta_i - \psi_i}}{\sum_{j=1}^{d} p_j e^{\theta_j - \psi_j}} \right)^{y_i} . \end{aligned}$$

We simplify the above by choosing p to be the vector with all components $1/d$ which corresponds to $\psi = 0$ so that $p_i e^{-\psi_i} = 1$. We will also choose $c(\psi) = n \log(d)$, so that

$$c(\theta) = n \log \left(\sum_{i=1}^{d} e^{\theta_i} \right),$$

Thus,

$$f_\theta(y) = \binom{n}{y} \prod_{i=1}^{d} \left(\frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}} \right)^{y_i},$$

and this is the PMF of the multinomial distribution with sample size n and probability vector having components

$$p_i(\theta) = \frac{e^{\theta_i}}{\sum_{j=1}^{d} e^{\theta_j}}.$$

This is not an identifiable parameterization.

The components of y sum to n so Y is concentrated on a hyperplane to which the vector $(1, 1, \cdots, 1)^T$ is perpendicular.

Eliminating a component of Y to get an identifiability would destroy symmetry of formulas and make everything harder and messier.