Excerpt from ”Exponential Family Notes”

Daniel J. Eck

August 6, 2020

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

$$f_\theta(y) = h(y) \exp\left(\langle y, \theta \rangle - c(\theta)\right) \tag{1}$$

and the word “density” here can refer to a probability mass function (PMF) or a probability density function (PDF) or to a probability mass-density function (PMDF) if we are referring to a distribution that is partly discrete and partly continuous (either some components of the Y are discrete and some continuous or some components are a mixture of discrete and continuous) or to a density with respect to an arbitrary positive measure in the sense of probability theory. The $h(y)$ arises from any term not containing the parameter that is dropped in going from log densities to log likelihood.

In this document we will discuss geometric properties of exponential families. We will write the density (1) as arising with respect to a generating measure $\lambda(dy) = h(y)dy$. Thus, the density for the exponential family is now

$$f_\theta(y) = \exp\left(\langle y, \theta \rangle - c(\theta)\right)$$

where the cumulant function $c(\theta)$ is the is log Laplace transformation of the measure λ

$$c(\theta) = \log\left(\int e^{y,\theta}\lambda(dy)\right).$$

The exponential family log likelihood takes the form

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

Identifiability and directions of constancy

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 (1) of one with respect to the other is equal to one with probability one. This happens if $Y'(\theta - \psi)$ 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

$$H = \{y : y'v = a\} \tag{2}$$

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

$$c(\theta + sv) = \log\left(\int e^{\langle y, \theta + sv \rangle} \lambda(dy)\right) = sa + \log\left(\int e^{\langle y, \theta \rangle} \lambda(dy)\right) = sa + c(\theta),$$

which immediately implies that

$$\begin{aligned} l(\theta + sv) &= \langle Y.\theta + sv \rangle - c(\theta + sv) \\ &= \langle Y.\theta \rangle + s\langle Y, v \rangle - (sa + c(\theta)) \\ &= \langle Y.\theta \rangle + sa - (sa + c(\theta)) \\ &= l(\theta). \end{aligned}$$

Therefore, we see that the canonical parameter vectors θ and $\theta + sv$ correspond to the same exponential family with probability equal to one for all $\theta \in \Theta$ when the canonical statistic is concentrated on a hyperplane (2). We summarize this as follows.

Theorem 1. *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 (8) 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 $Y'v$ 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.

When there exists a hyperplane H such that Y is concentrated to H, then there exists a non-zero direction such that Y exhibits no variability. Therefore, the variance matrix for Y is not full rank and the model is not identifiable. This idea can repeat itself. There can be several hyperplanes, nested within each other, such that the Y is concentrated on each of these hyperplanes. This can continue until we have exhausted the number of dimensions in Y. When this occurs, the model is completely generate. The variance matrix is 0 and every parameter value $\theta \in \Theta$ corresponds to the model that generates the observed data with probability one. In other words, all $\theta \in \Theta$ correspond to the same distribution.