Exponential Family Notes

Daniel J. Eck

Exponential families

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): The Binomial distribution Bin(n, p) can be parameterized as an exponential distribution parameterized as (1). We can write

$$\begin{aligned} l(p) &= \log\left(\binom{n}{y}\right) + y\log(p) + (n-y)\log(1-p) \\ &\propto y\log\left(\frac{p}{1-p}\right) + n\log(1-p) \\ &= y\theta - n\log\left(1+e^{\theta}\right) \\ &= \langle y, \theta \rangle - c(\theta), \end{aligned}$$

where

$\theta = \log\left(\frac{p}{1-p}\right)$ and $p = \frac{e^{\theta}}{1+e^{\theta}}$

$c(\theta) = n\log\left(1+e^{\theta}\right)$

$\log\left(\binom{n}{y}\right)$ is a function of the statistic y that is dropped because it does not contain the parameter. □

Densities

With our definitions we have some trouble writing down densities. First y is not the data; rather it is a statistic, a function of the data. 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 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 $h(w)$ arises from any term not containing the parameter that is dropped in going from log densities to log likelihood. We saw this in our Binomial distribution derivation.

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): The Binomial distribution Bin(n, p) can be parameterized as an exponential distribution. We can write

$$\begin{aligned}
f_p(y) &= \binom{n}{y} p^y (1-p)^{n-y} \\
&= h(y) \exp\left(y \log(p) + (n-y) \log(p)\right) \\
&= h(y) \exp\left(y \log\left(\frac{p}{1-p}\right) + n \log(p)\right) \\
&= h(y) \exp\left(y\theta - n \log\left(1 + e^\theta\right)\right) \\
&= h(y) \exp\left(\langle y, \theta \rangle - c(\theta)\right) \\
&= f_\theta(y)
\end{aligned}$$

where

$$\theta = \log\left(\tfrac{p}{1-p}\right) \text{ and } p = \tfrac{e^\theta}{1+e^\theta}$$

$$c(\theta) = n \log\left(1 + e^\theta\right)$$

$$h(y) = \tbinom{n}{y}$$

□

Example (Normal distribution): The Normal distribution $N(\mu, \sigma^2)$ can be parameterized as an exponential distribution. We can write

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

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{w^2 - 2w\mu + \mu^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2\sigma^2} + \frac{w\mu}{\sigma^2} - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(\sigma^2)\right)\right) \\
&= h(w) \exp\left(\langle Y(w), \theta\rangle - c(\theta)\right) \\
&= f_\theta(w)
\end{aligned}
$$

where

$h(w) = \frac{1}{\sqrt{2\pi}}$

$Y(w) = (w,\ -w^2)^T$ and $\theta = \left(\frac{\mu}{\sigma^2},\ \frac{1}{2\sigma^2}\right)^T$

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

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

$$
f_{\theta;\psi}(w) = e^{\langle Y(w), \theta - \psi\rangle - c(\theta) + c(\psi)} \tag{3}
$$

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). For any w such that $h(w) = 0$ (3) still makes sense because such w are not in the support of the distribution with parameter value ψ and hence do not not contribute to any probability or expectation calculation, so it does not matter how (3) is defined for such w. Now, since (3) is everywhere strictly positive, we see that every distribution in the family has the same support.

Cumulant Functions

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

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

Rearranging the above implies that

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

Being the expectation of a strictly positive quantity, the expectation here must always be strictly positive, so the logarithm is well-defined. By convention, for θ such that the expectation does not exist, we say $c(\theta) = \infty$.

In probability theory the cumulant function is the log Laplace transformation corresponding to the *generating measure* of the exponential family which is 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

$$\Theta = \{\theta : c(\theta) < \infty\}. \tag{4}$$

Then (1) and (2) define a distribution for all $\theta \in \Theta$, thus giving a statistical model that may be larger than the originally given model. We say an exponential family is *full* if its canonical parameter space is (4). Many commonly used statistical models are full exponential families. There is literature about so-called "curved exponential families" and other non-full exponential families, but we will not discuss them.

Moment and cumulant 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. This distinction is under the hood and not that important. The moment generating function of the canonical statistic, if it exists, is

given by

$$\begin{aligned} M_\theta(t) &= \mathrm{E}_\theta\left(e^{\langle Y,t\rangle}\right) \\ &= \int e^{\langle y,t\rangle} h(y) e^{(\langle y,\theta\rangle - c(\theta))} dy \\ &= \int h(y) e^{(\langle y,t+\theta\rangle - c(\theta))} dy \\ &= \int h(y) e^{(\langle y,t+\theta\rangle - c(\theta) \pm c(\theta+t))} dy \\ &= e^{c(\theta+t)-c(\theta)}. \end{aligned} \tag{5}$$

The moment generating function exists if it is finite on a neighborhood of zero, that is, if θ is an interior point of the full canonical parameter space (4). For other θ we say the moment generating function does not exist.

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,

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

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

$$k_\theta(t) = c(t+\theta) - c(\theta), \tag{6}$$

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

$$\begin{aligned} \nabla c(\theta) &= \mathrm{E}_\theta(Y) \\ \nabla^2 c(\theta) &= \mathrm{E}_\theta(YY^T) - [\mathrm{E}_\theta(Y)]\,[\mathrm{E}_\theta(Y)]^T = \mathrm{Var}_\theta(Y). \end{aligned} \tag{7}$$

In short, the mean and variance of the natural statistic always exist when θ is in the interior of the full canonical parameter space Θ, and they are given by derivatives of the cumulant function.

Verify that (7) **holds for the binomial, Poisson, and normal distriburions.**

Regular exponential families

This property of having mean and variance of the canonical statistic given by derivatives of the cumulant function is so nice that families which have it for all θ are given a special name. An

exponential family is regular if its full canonical parameter space (4) is an open set so that the moment and cumulant generating functions exist for all θ and 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. To see this, the success probability p has the parameter space $0 < p < 1$. 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. □

Identifiability and directions of constancy

In this section we will discuss geometric properties of exponential families. 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

$$H = \{y : y^T v = a\} \tag{8}$$

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 (8). 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 $Y^T v$ 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 the multinomial distribution is an exponential family and the usual vector statistic is canonical. To see this, let canonical parameter value ψ correspond to the multinomial distribution with sample size n and usual parameter vector p, and we find the exponential family generated by this distribution. Let d denote the dimension of y and θ, let $\binom{n}{y}$ denote multinomial coefficients, and let S denote the sample space of the multinomial distribution (vectors having nonnegative integer components that sum to n).

In the same vein as (3), 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

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

where the last equality follows from the multinomial theorem. Then (3) gives

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

$$= \binom{n}{y} \prod_{i=1}^{d} \left(\frac{p_i e^{\theta_i - \psi_i}}{\sum_{j=1}^{d} p_j e^{\theta_j - \psi_j}} \right)^{y_i}.$$

We simplify the above by choosing p and ψ so that $p_i e^{-\psi_i} = 1$ for all i and $c(\psi) = 0$, so

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

and

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

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, however, 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, hence by Theorem 1 a direction of constancy of the family. Eliminating a component of Y to get an identifiability would destroy symmetry of formulas and make everything harder and messier. Best to wait until when (if ever) identifiability becomes absolutely necessary. □

The Right Way[1] (IMHO) to deal with nonidentifiability, which is also called collinearity in the regression context, is the way the `R` functions `lm` and `glm` deal with it. (We will have to see how linear and generalized linear models relate to exponential families before this becomes fully clear, but I assure you this is how what they do relates to a general exponential family). When you find you have a non-identifiable parameterization, you have $Y^T v$ constant with probability one. Pick any i such that $v_i \neq 0$ and fix $\theta_i = 0$ giving a submodel that (we claim) has all the distributions of the original one (we have to show this).

For any parameter vector θ in the original model (with θ_i free to vary) we know that $\theta + sv$ corresponds to the same distribution for all s. Choose s such that $\theta_i + sv_i = 0$, which is possible because $v_i \neq 0$, hence we see that this distribution is in the new family obtained by constraining θ_i to be zero (and the other components of θ vary freely).

This new model obtained by setting θ_i equal to zero is another exponential family. Its canonical statistic and parameter are just those of the original family with the i-th component eliminated. Its cumulant function is just that of the original family with the i-th component of the parameter set to zero. This new model need not be identifiable, but if not there is another direction of constancy and the process can be repeated until identifiability is achieved (which it must because the dimension

[1]The Right Way is borrowed vernacular from Charles Geyer. The Right Way means anything that is not obviously the Wrong Way. There can be several Right Ways, and choosing among them can be subjective.

of the sample space and parameter space decreases in each step and cannot go below zero, and if it gets to zero the canonical statistic is concentrated at a single point, hence there is only one distribution in the family, and identifiability vacuously holds).

This is what `lm` and `glm` do. If there is non-identifiability (collinearity), they report `NA` for some regression coefficients. This means that the corresponding predictors have been "dropped" but this is equivalent to saying that the regression coefficients reported to be `NA` have actually been constrained to be equal to zero.

Mean Value Parameterization

The mean of the canonical statistic $\mathrm{E}_\theta(Y)$ is also a parameter. It is given as a function of the canonical parameter θ,

$$\mu = \mathrm{E}_\theta(Y) = \nabla c(\theta) = g(\theta). \tag{10}$$

We will refer to $g(\theta)$ as the change-of-parameter map (or change-of-parameter) from canonical parameter θ to mean value parameter μ.

Theorem 2. *For a regular full exponential family, the change-of-parameter from canonical to mean value parameter is invertible if the model is identifiable. Moreover both the change-of-parameter and its inverse are infinitely differentiable.*

To prove this let μ be a possible value of the mean value parameter (that is, $\mu = g(\theta)$ for some θ) and consider the function

$$h(\theta) = \langle \mu, \theta \rangle - c(\theta). \tag{11}$$

The second derivative of h is $-\nabla^2 c(\theta)$ which is equal to $-\operatorname{Var}_\theta(Y)$, and this is a negative definite matrix (**Why?**) Hence (11) is a strictly concave function by Theorem 2.14 in Rockafellar and Wets [1998], and this implies that the maximum of (11) is unique if it exists by Theorem 2.6 in Rockafellar and Wets [1998]. Moreover, we know a solution exists because the derivative of (11) is $\nabla h(\theta) = \mu - \nabla c(\theta)$, and we specified that $\mu = \nabla c(\theta)$ for some θ.

Show that cumulant functions are infinitely differentiable and are therefore continuously differentiable. Now we see that the Jacobian matrix of the change-of-parameter is

$$\nabla g(\theta) = \nabla^2 c(\theta)$$

which we (you) have already shown is nonsingular. The inverse function theorem (Browder, 1996, Theorems 8.15 and 8.27) thus says that g is locally invertible, and the local inverse must agree with the global inverse which we have already shown exists. The inverse function theorem goes on to state that the derivative of the inverse is the inverse of the derivative

$$\nabla g^{-1}(\theta) = \left[\nabla g(\theta)\right]^{-1}, \qquad \text{when } \mu = g(\theta) \text{ and } \theta = g^{-1}(\mu).$$

Now show that $g^{-1}(\theta)$ is infinitely differentiable.

Multivariate Monotonicity

A mapping from $g : \mathbb{R}^d \to \mathbb{R}^d$ is multivariate monotone (Rockafellar and Wets, 1998, Definition 12.1) if

$$\left[g(x_1) - g(x_2)\right]^T (x_1 - x_2) \geq 0, \qquad \text{for } x_1 \text{ and } x_2 \in \mathbb{R}^d, \tag{12}$$

and strictly multivariate monotone if (12) holds with strict inequality whenever $x_1 \neq x_2$. If g is differentiable, then by Proposition 12.3 in Rockafellar and Wets [1998] it is multivariate monotone if and only if the symmetric part of the Jacobian matrix ∇g is positive-semidefinite for each x. A sufficient but not necessary condition for g to be strictly multivariate monotone is that the symmetric part of ∇g be positive definite for each x.

Let g be the change-of-parameters mapping from canonical to mean value parameters (10) then we showed in the previous section that its Jacobian matrix is positive semidefinite in general and strictly positive definite when the model is identifiable. Thus this change-of-parameter is multivariate monotone in general and strictly multivariate monotone when the model is identifiable.

Thus, if μ_1 corresponds to θ_1 and μ_2 to θ_2, we have

$$(\mu_1 - \mu_2)^T(\theta_1 - \theta_2) > 0, \qquad \text{whenever;} \theta_1 \neq \theta_2. \tag{13}$$

In general, this is all we can say about the map from canonical to mean value parameters. However, there is a casual version of (13) which eases interpretation. If we rewrite (13) using subscripts

$$\sum_{i=1}^{d} (\mu_{1i} - \mu_{2i})(\theta_{1i} - \theta_{2i}) > 0$$

and consider θ_1 and θ_2 that differ in only one coordinate, say the kth, then we get

$$(\mu_{1k} - \mu_{2k})(\theta_{1k} - \theta_{2k}) > 0,$$

which says *if we increase one component of the canonical parameter vector, leaving the other components fixed, then the corresponding component of the mean value parameter vector also increases, and the other components can go any which way*. This is easier to explain than the full multivariate monotonicity property, but is not equivalent to it. The casual property is not enough to make some arguments about exponential families that are needed in applications (for example in the Appendix of Shaw and Geyer, 2010).

Here is another rewrite of (13) that preserves its full force. Fix a vector $v \neq 0$. Write $\theta_2 = \theta$ and $\theta_1 = \theta + sv$, so multivariate monotonicity (12) becomes

$$\left[g(\theta + sv) - g(\theta)\right]^T v > 0, \qquad \text{for } s \neq 0.$$

Differentiate with respect to s and set $s = 0$, which gives the so-called directional derivative of g in the direction v at the point θ

$$\nabla g(\theta; v) = v^T \left[\nabla g(\theta)\right] v = v^T \left[\nabla^2 c(\theta)\right] v. \tag{14}$$

We know that $\nabla^2 c(\theta)$ is positive semi-definite in general and strictly positive definite when the model is identifiable. Hence we see (again) that the θ to μ mapping is multivariate monotone in general and strictly multivariate monotone when the model is identifiable.

Partial derivatives are special cases of directional derivatives when the vector v points along a coordinate direction (only one component of v is nonzero). So the casual property only says that all the partial derivatives are nonzero and this corresponds to asserting (14) with v being along coordinate directions, and this is equivalent to asserting that the diagonal components of $\nabla^2 c(\theta)$ are positive. And now we clearly see how the casual property is indeed casual. It only asserts that the diagonal elements of $\nabla^2 c(\theta)$ are positive, which is far from implying that $\nabla^2 c(\theta)$ is a positive definite matrix.

Maximum likelihood estimation

The derivative of the log likelihood is

$$\nabla l(\theta) = y - \nabla c(\theta).$$

The second derivative is

$$\nabla^2 l(\theta) = -\nabla^2 c(\theta).$$

Hence observed and expected Fisher information for the canonical parameter vector θ are the same

$$I(\theta) = \nabla^2 c(\theta). \tag{15}$$

Fisher information measures the expected curvature of the log likelihood around the true parameter value. If the likelihood is sharply curved around θ – the expected information $I(\theta)$ is large – then a small change in θ can lead to a drastic decrease in the likelihood. Conversely, if $I(\theta)$ is small then small changes in θ will not affect the likelihood that much. These heuristics are important when we cover separation and non-identifiability.

When the model is identifiable, the canonical statistic vector Y is not concentrated on a hyperplane, the second derivative is negative definite everywhere, hence the log likelihood is strictly concave, hence the maximum likelihood estimate is unique if it exists. Thus,

$$\begin{aligned} y &= \nabla c(\hat{\theta}), \\ \hat{\theta} &= g^{-1}(y). \end{aligned}$$

Derive the MLEs of the canonical parameters of the binomial, Poisson, and normal distributions.

Non-Existence of the MLE

Unlike our proof of Theorem 2 where we assumed the existence of a solution, we cannot prove the maximum likelihood estimate (for the canonical parameter) exists. Consider the binomial distribution. The MLE for the usual parameterization is $\hat{p} = y/n$. The canonical parameter is $\theta = \text{logit}(p)$. But $\hat{\theta} = \text{logit}(\hat{p})$ does not exist when $\hat{p} = 0$ or $\hat{p} = 1$, which is when we observe zero successes or when we observe n successes in n trials.

One might think the lesson to draw from this is not to use the canonical parameterization, but it turns out that generalized linear models and log-linear models for categorical data and other applications of exponential families always use the canonical parameterization for many reasons. Hence we have to deal with possible non-existence of the MLE. We will revisit this topic when we discuss GLMs.

Observed equals expected

For a regular full exponential family, the MLE cannot be on the boundary of the canonical parameter space (regular means the boundary is empty), and the MLE, if it exists, must be a point where the first derivative is zero, that is, a θ that satisfies

$$y = \nabla c(\theta) = \mathrm{E}_\theta(Y).$$

The MLE is the (unique if the model is identifiable) parameter value that makes the observed value of the canonical statistic equal to its expected value. We call this the ***observed equals expected*** property of maximum likelihood in exponential families. This property is even simpler to express in terms of the mean value parameter. By invariance of maximum likelihood under change-of-parameter, the MLE for μ is

$$\hat{\mu} = \nabla c(\hat{\theta})$$

and the observed equals expected property is therefore

$$y = \hat{\mu}. \tag{16}$$

Independent and Identically Distributed

Suppose $y_1, \ldots, y_n$ are independent and identically distributed (IID) from some distribution in an exponential family (unlike our notation in the preceding section, y_i are not components of the canonical statistic vector but rather IID realizations of the canonical statistic vector, so each y_i is a vector). The log likelihood for sample size n is

$$l_n(\theta) = \sum_{i=1}^n \left[\langle y_i, \theta \rangle - c(\theta) \right] = \langle \sum_{i=1}^n y_i, \theta \rangle - n c(\theta), \tag{17}$$

and we see that the above log likelihood is an exponential family with

- canonical statistic $\sum_{i=1}^n y_i$,
- cumulant function $\theta \mapsto nc(\theta)$, and
- canonical parameter θ and full canonical parameter space Θ the same as the originally given family from which every observation is a member.

Thus IID sampling gives us a new exponential family, but still an exponential family.

Asymptotics of Maximum Likelihood

Rewrite (17) as

$$l_n(\theta) = n \left[\langle \bar{y}_n, \theta \rangle - c(\theta) \right]$$

so that

$$\nabla l_n(\theta) = n \left[\bar{y}_n - \nabla c(\theta) \right].$$

From which we see that for an identifiable regular full exponential family where the MLE must be a point where the first derivative is zero, we can write

$$\begin{aligned} \nabla l_n(\theta) = n \left[\bar{y}_n - \nabla c(\theta) \right] &= 0, \\ \bar{y}_n - \nabla c(\hat{\theta}) &= 0, \\ \bar{y}_n &= \nabla c(\hat{\theta}). \end{aligned}$$

Recall the change-of-parameters mapping $g : \theta \mapsto \mu$ given by (10) in the mean value parameters section. We can write

$$\hat{\theta}_n = g^{-1}(\bar{y}_n). \tag{18}$$

More precisely, (18) holds when the MLE exists (when the MLE does not exist, $\bar{y}_n$ is not in the domain of g^{-1}, which is in the range of g).

By the multivariate central limit theorem (CLT)

$$\sqrt{n} \left(\bar{y}_n - \mu \right) \to N \left(0, I(\theta) \right)$$

and we know that g^{-1} is differentiable (Theorem 2) with the derivative given by

$$\nabla g^{-1}(\theta) = \left[\nabla g(\theta) \right]^{-1}, \qquad \text{where } \mu = g(\theta) \text{ and } \theta = g^{-1}(\mu).$$

So the usual asymptotics of maximum likelihood

$$\sqrt{n} \left(\hat{\theta}_n - \theta \right) \to N \left(0, I(\theta)^{-1} \right) \tag{19}$$

is just the multivariate delta method applied to the multivariate CLT.

Finite sample concentration of MLE

The previous section is devoted to large sample properties of maximum likelihood estimation within the context of regular full exponential families. These properties are especially relevant for statistical inference. MLEs of parameters in regular full exponential families also possess desirable finite sample properties. We first motivate the concept of sub-Gaussian and sub-exponential random variables which represent classes of desirable tail behavior for statistical models. The following definitions come from Wainwright [2019]:

Definition 1. *A random variable Y with mean $\mu = \mathrm{E}(Y)$ is* sub-Gaussian *if there exists a positive number λ such that*

$$E\left(e^{\phi(Y-\mu)}\right) \leq e^{\lambda^2\phi^2/2} \qquad \textit{for all } \phi \in \mathbb{R}.$$

Definition 2. *A random variable Y with mean $\mu = \mathrm{E}(Y)$ is* sub-exponential *if there exist non-negative numbers (λ, b) such that*

$$E\left(e^{\phi(Y-\mu)}\right) \leq e^{\lambda^2\phi^2/2} \qquad \textit{for all } |\phi| < 1/b.$$

Let Y be a univariate regular full exponential family with canonical parameter θ and **show that Y is sub-exponential**. Now let $\hat{\theta}$ be the MLE for the canonical parameter θ. We now show that the MLE of an exponential family obeys sub-exponential concentration. Consider a Taylor expansion of the score function of an exponential family evaluated at the MLE

$$\begin{aligned} 0 = \nabla l_n(\hat{\theta}) &= \nabla l_n(\theta) + \nabla^2 l_n(\theta)(\hat{\theta} - \theta) + R_n \\ &= \sum_{i=1}^{n} \{y_i - \nabla c(\theta)\} + \nabla^2 l_n(\theta)(\hat{\theta} - \theta) + R_n, \end{aligned}$$

where $R_n = o_P(n^{-1/2})$. Notice that $\sum_{i=1}^{n}\{y_i - \nabla c(\theta)\}$ is a sum of mean zero sub-exponential random variables, and is also sub-exponential [Wainwright, 2019, Chapter 2]. Furthermore, the negative and/or a scalar products of $\sum_{i=1}^{n}\{y_i - \nabla c(\theta)\}$ are also sub-exponential. After rearranging terms in the above displayed equation we see that

$$(\hat{\theta} - \theta) = -\{\nabla^2 l_n(\theta)\}^{-1} \sum_{i=1}^{n} \{y_i - \nabla c(\theta)\} + \widetilde{R}_n,$$

where $\widetilde{R}_n = -\{\nabla^2 l_n(\theta)\}^{-1} R_n = o_P(n^{-1/2})$. Putting all of this together yields

$$\mathbb{P}\left((\hat{\theta} - \theta) \geq t\right) = \mathbb{P}\left(-\{\nabla^2 l_n(\theta)\}^{-1} \sum_{i=1}^{n} \{y_i - \nabla c(\theta)\} \geq t - \widetilde{R}_n\right),$$

where $t > 0$. Handwaving slightly, there exists a number $A > 0$ such that, for n large,

$$\mathbb{P}\left(-\{\nabla^2 l_n(\theta)\}^{-1} \sum_{i=1}^{n} \{y_i - \nabla c(\theta)\} \geq t - \widetilde{R}_n\right)$$

$$= \mathbb{P}\left(-\{n^{-1}\nabla^2 l_n(\theta)\}^{-1}\left(n^{-1}\sum_{i=1}^{n}\{y_i - \nabla c(\theta)\}\right) \geq t - \widetilde{R}_n\right)$$
$$\leq \mathbb{P}\left(-An^{-1}\sum_{i=1}^{n}\{y_i - \nabla c(\theta)\} \geq t\right).$$

Page 29 in Wainwright [2019] implies that, for the sub-exponential variable $-An^{-1}\sum_{i=1}^{n}\{y_i - \nabla c(\theta)\}$, we have parameters (λ, b) such that

$$\mathbb{P}\left(-An^{-1}\sum_{i=1}^{n}\{y_i - \nabla c(\theta)\} \geq t\right) \leq \begin{cases} \exp\left(-\frac{nA^2t^2}{2\lambda^2}\right), & \text{for } 0 \leq t \leq \frac{\lambda^2}{nAb}, \\ \exp\left(-\frac{nAt}{2b}\right), & \text{for } t > \frac{\lambda^2}{nAb}. \end{cases}$$

We can therefore conclude that the MLE of θ exhibits sub-exponential concentration following the logic that $(\hat{\theta} - \theta)$ has the same tail bounds as a sub-exponential random variable. We can use these results to obtain the rate of convergence

$$\mathbb{P}\left((\hat{\theta} - \theta) \geq \frac{\log(n)}{n}\right) = O\left(n^{-\frac{A}{2b}}\right),$$

where $t = \log(n)/n$.

Canonical linear submodels: intro to GLMs

We now motivate generalized linear models (GLMs) within the context of exponential theory. We will not yet discuss specific examples of GLMs or extensions beyond canonical representations. We first present canonical affine submodels of an exponential family. A canonical affine submodel of an exponential family is a submodel having parameterization

$$\theta = a + M\beta$$

where $\theta \in \mathbb{R}^n$ is the canonical parameter vector corresponding to the original exponential family, $\beta \in \mathbb{R}^p$ is the canonical parameter vector for the submodel, a is known offset vector, and M is a known matrix. The matrix M is usually called the *model matrix* in the terminology used by the `R` functions `lm` and `glm`. The vector a is called the offset vector in the terminology used by the `R` functions `lm` and `glm`. In most applications the offset vector is not used giving parameterization

$$\theta = M\beta,$$

in which case we say the submodel is a *canonical linear submodel*. We will restrict attention to the canonical linear submodel. The canonical linear submodel log likelihood is given by

$$\begin{aligned} l(\theta) &= \langle y, \theta\rangle - c(\theta) \\ &= \langle y, M\beta\rangle - c(M\beta) \\ &= \langle M^T y, \beta\rangle - c(M\beta), \end{aligned} \tag{20}$$

and we see that we again have an exponential family with

- canonical statistic $M^T y$,
- cumulant function $\beta \mapsto c(M\beta)$, and
- submodel canonical parameter vector β.

If θ varies freely (over a whole vector space), then β also varies freely (over a whole vector space of lower dimension, $p \le n$). But if the originally given full canonical parameter space was Θ, then the full submodel canonical parameter space is

$$B = \{\beta : M\beta \in \Theta\}.$$

Thus a canonical linear submodel gives us a new exponential family, with lower-dimensional canonical parameter and statistic. The submodel exponential family is full if the original exponential family was full. Notice that θ and B are defined through the column space of M, not the particular model matrix M, the particular β value does not determine the submodel uniquely.

To distinguish between the submodel and the originally given exponential family, we often call the latter the *saturated model*. Now we have four parameters: the saturated model canonical and mean value parameters θ and μ and the canonical linear submodel canonical and mean value parameters β and $\tau = M^T\mu$. Relations between these parameterizations are given in Figure 1.

The observed equals expected property for the submodel is

$$\hat{\tau} = M^T\hat{\mu} = M^T y. \tag{21}$$

We cannot actually solve these equations for $\hat{\mu}$ because the mapping $\mu \to M^T\mu$ is usually not one-to-one (the $n > p$ case where $M \in \mathbb{R}^{n\times p}$ and is full column rank). Hence we cannot determine $\hat{\theta}$ and $\hat{\beta}$ from them either. The only way to determine the MLE is to maximize the log likelihood (20) for β to obtain $\hat{\beta}$ and then $\hat{\theta} = M\hat{\beta}$ and $\hat{\mu} = \nabla c(\hat{\theta})$ and $\hat{\tau} = M^T\hat{\mu}$. But the observed equals expected property is nevertheless very important.

Derive the asymptotic distribution for the MLE of $\hat{\beta}$ and $\hat{\tau}$.

Recall that the saturated model canonical parameter vector θ is "linked" to the saturated model mean value parameter vector through the change-of-parameter mappings $g(\theta)$. We can reparameterize $\theta = M\beta$ and write

$$\mathrm{E}_\theta(Y) = \mu = g(M\beta)$$

which implies that we can write

$$g^{-1}\left(\mathrm{E}_\theta(Y)\right) = M\beta.$$

Therefore, a linear function of the canonical submodel parameter vector is linked to the mean of the exponential family through the inverse change-of-parameter mapping g^{-1}. This is the basis of

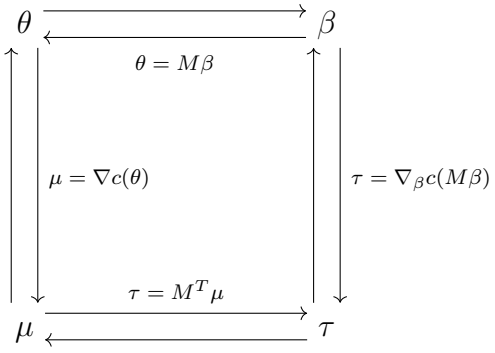

Figure 1: A depiction of the transformations necessary to change between parameterizations. Arrows going in opposite directions specify transformations and their inverses. M is a known model matrix of full column rank, and c is the cumulant function for the exponential family model.

exponential family generalized linear models with link function g^{-1}. Note that most treatments of GLMs will present g as the link function. Instead we motivated g as the change of parameters mapping from canonical to mean-value parameters.

Inference

Recall that the canonical submodel log likelihood (20) takes the form

$$l(\beta) = \langle M^T y, \beta \rangle - c(M\beta)$$

when the offset is ignored. From invariance of maximum likelihood estimation derived in (19) or direct arguments, we have that the asymptotic distribution of the MLE $\hat{\beta}$ takes the form

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} N(0, \Sigma^{-1})$$

where $\Sigma = \mathrm{E}\left(-\nabla^2 l(\beta)\right)$ is the Fisher information matrix corresponding to the canonical linear submodel.

Wald inference

Let $\widehat{\Sigma}$ be estimated Σ using the MLE $\hat{\beta}$ in place of β. In particular, the jth element $\hat{\beta}_j$ of $\hat{\beta}$ is asymptotically normal with asymptotic variance

$$\widehat{\mathrm{Var}}(\hat{\beta}_j) = jth \text{ diagonal element of } \widehat{\Sigma}.$$

The Wald Z statistic for testing $H_o : \beta_j = \beta_{jo}$ is

$$z_W = \frac{\hat{\beta}_j - \beta_{jo}}{\text{se}(\hat{\beta}_j)} \quad \overset{H_o}{\sim} \quad N(0,1)$$

where $\text{se}(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}$. We can construct $(1-\alpha) \times 100\%$ Wald based confidence intervals of the form

$$\hat{\beta}_j \pm z_{\alpha/2}\text{se}(\hat{\beta}_j)$$

for some error tolerance $0 < \alpha < 1$.

Deviance, Goodness of Fit, and likelihood ratios

To motivate the deviance of a statistical model we will revisit the mean-value parameter vector μ and rewrite the log likelihood in the notation of this parameterization as $l(\mu; y)$. From the observed equals expected property we have that the unrestricted MLE of μ is y. Now consider a canonical submodel (GLM) of the form $\theta = M\beta$. Let $\hat{\mu}$ be the MLE of μ restricted to an identifiable canonical submodel ($\hat{\mu} = \nabla c(\hat{\theta})$ where $\hat{\theta} = M\hat{\beta}$). It follows that

$$l(y; y) \geq l(\hat{\mu}; y).$$

We refer to the unrestricted case, in which each observation has its own mean ($\hat{\mu} = y$), is called the *saturated model*.

With all of this in mind, the *deviance* of the GLM is

$$D(y; \hat{\mu}) = -2\left(l(\hat{\mu}; y) - l(y; y)\right).$$

We see that the deviance statistic is a function of a ratio of two likelihoods, one corresponding to the canonical submodel and one corresponding to the saturated model. The deviance statistic has approximate distribution

$$D(y; \hat{\mu}) \quad \overset{H_o}{\sim} \quad \chi^2_{\text{df}}$$

where H_o is that the canonical submodel is correct and the alternative test is that the canonical submodel is incorrect but the saturated model is correct, df = $n - p$, n is the sample size, and p is the number of parameters in the canonical submodel. So we reject correctness of the canonical submodel when

$$D(y; \hat{\mu}) \quad > \quad \chi^2_{\text{df}}(\alpha).$$

(Note that the χ^2 approximation can be poor.)

We can use deviance based testing to nested models. Let $\mathcal{M}_0$ and $\mathcal{M}_1$ both be canonical submodels. We say that $\mathcal{M}_0$ is *nested* within $\mathcal{M}_1$ when every distribution in $\mathcal{M}_0$ is also in $\mathcal{M}_1$ but

not vice-versa. That is, μ is more restricted under $\mathcal{M}_0$ than under $\mathcal{M}_1$. Let $\hat{\mu}_0$ be the MLE of μ under $\mathcal{M}_0$, and let $\hat{\mu}_1$ be the MLE of μ under $\mathcal{M}_1$. We can use this framework for testing

$$H_0 : \mathcal{M}_0 \text{ true} \qquad H_a : \mathcal{M}_1 \text{ true, but not } \mathcal{M}_0$$

using the likelihood ratio χ^2 statistic given by

$$\begin{aligned} -2\left[l(\hat{\mu}_0; y) - l(\hat{\mu}_1; y)\right] &= -2\left[l(\hat{\mu}_0; y) - l(y; y)\right] - \{-2\left[l(\hat{\mu}_1; y) - l(y; y)\right]\} \\ &= D(y; \hat{\mu}_0) - D(y; \hat{\mu}_1) \\ &\approx \chi^2_{\text{df}}, \end{aligned}$$

where df = $p_1 - p_0$, p_1 is the number of parameters in $\mathcal{M}_1$, and p_0 is the number of parameters in $\mathcal{M}_0$. (Note that the χ^2 approximation is often adequate here even it isn't adequate for the saturated model provided that $\mathcal{M}_1$ is not too close to saturated.)

Optimization

In this section we discuss optimization routines for estimating parameters in canonical exponential family linear submodels. The goal will be to find

$$\text{argmax}_\beta l(\beta) \quad = \quad \text{argmax}_\beta \left[\langle M^T y, \beta\rangle - c(M\beta)\right] \tag{22}$$

in identifiable models. Note that we will blend our notation with the notation in Chapter 4 of Agresti [2013] when we define the Newton-Raphson, Fisher scoring, and IRLS algorithms.

Newton-Raphson method: A classic algorithm for handling iterative solutions of nonlinear systems of equations is the *Newton-Raphson algorithm*. This algorithm begins with an initial guess β_0 for the solution. It obtains a second guess by approximating the function to be maximized in a neighborhood of the initial guess by a second-degree polynomial and then finding the location of the polynomial's maximum value. This process is repeated iteratively until the discrepancy in successive evaluations of the objective function evaluate along the sequence of iterates is smaller than some convergence threshold. The sequence of iterates that this algorithm generates converge to a solution $\hat{\beta}$ when the optimization function is suitable (full rank properly conditioned Fisher Information matrix) and/or the initial guess is good.

We now explain the Newton-Raphson algorithm formally. Let $U(\beta) = \nabla l(\beta)$ be the score function corresponding to the log likelihood canonical linear exponential family submodel, and let $H(\beta) = \nabla^2 l(\beta)$ denote the Hessian matrix. At iteration k, consider the following second order Taylor series approximation of $l(\beta)$,

$$l(\beta) \approx l(\beta_k) + U(\beta_k)^T(\beta - \beta_k) + \frac{(\beta - \beta_k)^T H(\beta_k)(\beta - \beta_k)}{2}. \tag{23}$$

Now solving

$$U(\beta) \approx U(\beta_k) + H(\beta_k)(\beta - \beta_k) = 0$$

for β yields the next guess. That guess is

$$\beta_{k+1} = \beta_k - H(\beta_k)^{-1} U(\beta_k). \tag{24}$$

This algorithm is locally fast, exhibits quadratic convergence, provided that it converges. Convergence is likely in identifiable models where $H(\beta_0)$ is positive definite. However, the Newton-Raphson method can be quite sensitive to the choice of starting values β_0.

For many identifiable GLMs, including Poisson models with log link and binomial models with logit link, with full rank model matrices the Hessian is negative definite and the log likelihood is a strictly concave function. The maximum likelihood estimators of model parameters exist and are unique under quite general conditions [Wedderburn, 1976].

Fisher scoring algorithm: The *Fisher scoring algorithm* is an alternative method for solving systems of equations. It resembles the Newton-Raphson algorithm, the distinction being with the Hessian matrix used in the Newton updates. Fisher scoring uses the expected Fisher information matrix instead of the Hessian which is the observed Fisher information matrix.

We will let $\mathcal{H}$ be the expected information matrix so that $\mathcal{H}(\beta) = -\mathrm{E}\left\{\nabla^2 l(\beta)\right\}$. The Newton update step for the Fisher scoring method is

$$\beta_{k+1} = \beta_k + \left\{\mathcal{H}(\beta_k)\right\}^{-1} U(\beta_k).$$

or

$$\mathcal{H}(\beta_k)\beta_{k+1} = \mathcal{H}(\beta_k)\beta_k + U(\beta_k). \tag{25}$$

The chain rule and an appeal to Figure 1 allows us to write $\mathcal{H}(\beta) = M^T W(\beta) M$ where

$$W(\beta) = \left\{\nabla^2_\theta c(\theta^*)\right\}|_{\theta^* = M\beta}.$$

The estimated asymptotic covariance matrix $\mathcal{H}^{-1}$ of $\hat{\beta}$ occurs as a by-product of this algorithm as $\left\{\mathcal{H}(\beta_k)\right\}^{-1}$ where k is an iteration number at which convergence is deemed to have occurred.

For both Fisher scoring and Newton-Raphson, the score function $U(\beta)$ can be written as

$$U(\beta) = \nabla l(\beta) = M^T \left\{Y - \nabla_\theta c(\theta^*)|_{\theta^* = M\beta}\right\}.$$

For GLMs with canonical link (the entirety of these notes), we have that the observed and expected information are the same. This is a consequence of the observed equals expected property $\hat{\mu} = y$ that is noted above. For noncanonical link models (which we see later), Fisher scoring has the advantages that it produces the asymptotic covariance matrix as a by-product, the expected

information is necessarily nonnegative definite, and as seen next, it is closely related to weighted least-squares methods for ordinary linear models. However, it need have second-order convergence, and for complex models the observed information is often easier to calculate.

IRLS algorithm: A relation exists between *weighted least-squares estimation* and using the Fisher scoring algorithm to find MLEs. We refer here to the general linear model of the form

$$Z = M\beta + \varepsilon,$$

where V be the covariance matrix of ε. The weighted least-squares (WLS) estimator of β is

$$\hat{\beta}_{\text{WLS}} = \left(M^T V^{-1} M\right)^{-1} M^T V^{-1} Z.$$

We will now motivate the iterative (re)weighted least squares algorithm for estimation of the submodel canonical parameter vector β. Recall that $\mathcal{H}(\beta) = M^T W(\beta) M$ and note that we can write the score function $U(\beta)$ as

$$U(\beta) = M^T W(\beta) W^{-1}(\beta)(y - \mu(\beta)),$$

Now, noting the right hand side of (25),

$$\begin{aligned}
\mathcal{H}(\beta)\beta + U(\beta) &= (M^T W(\beta) M)\beta + M^T W(\beta) W^{-1}(\beta)(y - \mu(\beta)) \\
&= M^T W(\beta) \left[M\beta + W^{-1}(\beta)(y - \mu(\beta))\right] \\
&= M^T W(\beta) z(\beta).
\end{aligned}$$

Another referral to (25) gives

$$\mathcal{H}(\beta_k)\beta_{k+1} = M^T W(\beta_k) z(\beta_k) \qquad \Longrightarrow \qquad (M^T W(\beta_k) M)\beta_{k+1} = M^T W(\beta_k) z(\beta_k).$$

The above equation has a solution of the form

$$\beta_{k+1} = (M^T W(\beta_k) M)^{-1} M^T W(\beta_k) z(\beta_k),$$

where

- the "response" is $z(\beta) = M\beta + W^{-1}(\beta)(y - \mu(\beta))$
- $\mu(\beta)$ is the mean value parameter defined as a function of β

Note that although we follow a similar notation as Agresti [2013], our algorithm and our $W(\beta)$ quantity is slightly different than that in Agresti [2013].

Quasi-Newton methods: Quasi-Newton methods are modifications of (24), where the J matrix is approximated (by the secant method for example), typically because an explicit formula for J is

not available. One such Quasi-Newton method is the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm for solving unconstrained nonlinear optimization problems. We present the steps for the BFGS algorithm in our optimization context (22) where l is a differentiable scalar function.

The algorithm begins at a user-specified initial value (initial estimate for the optimal value) $\beta_{k=0}$ and an initial guess B_0. Then proceed iteratively to get a better estimate of β at each stage. The steps are:

1. Obtain a direction x_k by solving $B_k x_k = \nabla_\beta l(\beta_k)$.
2. Perform a one-dimensional optimization (line search) to find an acceptable stepsize α_k to be made in the direction x_k, so that $x_k = \text{arg min} - l(\beta_k + \alpha x_k)$.
3. Set $s_k = \alpha_k x_k$ and update $\beta_{k+1} = \beta_k + s_k$.
4. Set $y_k = \nabla_\beta(\beta_k) - \nabla_\beta(\beta_{k+1})$.
5. Update
$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k^T}{s_k^T B_k s_k}.$$

The BFGS algorithm is a possible optimization method in R's optim function. There is also a variant of BFGS in R called L-BFGS-B. This algorithm extends a limited memory variant of BFGS to handle bound constraints in the domain space of $l(\beta)$. Bound constraints are of the form $l_i \leq \beta_i \leq u_i$ where β_i is one component of the canonical parameter vector β.

Stochastic Gradient Decent (SGD): The SGD algorithm with averaging [Polyak and Juditsky, 1992] is a computationally fast, and scalable optimization method that is applicable in large scale applications with applications in online learning. We will consider SGD in our context for estimating $\hat{\beta}$ in exponential family models. The SGD algorithm only requires use of one (or relatively few) data points at each iteration, which makes it computationally superior to other optimization routines. Conventional implementations average over all iterations to derive the final estimator. Consistency and asymptotic normality properties of SGD have been verified in Polyak and Juditsky [1992] for both parameter estimation. In addition to these results, Rakhlin et al. [2011] derived the optimal rates of $O(1/n)$ for the objective function under smoothness and strong convexity assumptions. These assumptions hold in our modeling context.

We will suppose that we have data $D_n = Z_1, \ldots, Z_n$ drawn as iid copies of $Z = (y, x)$. In this case we assume that the both the response and the predictors are random from some common generative process. We will suppose that the conditional distibution of interest $y|x$ can then be parameterized as an exponential family with log likelihood (20) where the matrix M has rows x_i^T, $i = 1, \ldots, n$. The SGD algorithm updates as follows:

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \eta_k \nabla l(\hat{\beta}_{k-1}; Z_k)$$

for $k = 1, \ldots, n$ where η is called the learning rate. We set the learning rate to be $\eta_k = Ck^{-\alpha}$ for some $\alpha \in (0, 1)$. The final SGD estimators is then

$$\hat{\beta} = \frac{1}{n} \sum_{k=1}^{n} \hat{\beta}_k.$$

We will revisit optimization when we discuss specific GLMs.

Sufficiency

A (possibly vector-valued) statistic is *sufficient* if the conditional distribution of the full data given this statistic does not depend on the parameter.

The interpretation is that the full data provides no information about the parameter that is not already provided by the sufficient statistic. The principle of sufficiency follows: all inference should depend on the data only through sufficient statistics.

The Fisher-Neyman factorization criterion [Lehmann, 1959, Corollary 1 of Chapter 2] says that a statistic is sufficient if and only if the likelihood depends on the whole data only through that statistic.

Lemma 1. *The canonical statistic vector of an exponential family is a sufficient statistic.*

The proof of the above Lemma is left as an exercise for the reader.

Sufficient dimension reduction is a whole field of study. However, the *original* "sufficient dimension reduction" theory was about exponential families. The so-called Pitman-Koopman-Darmois theorem (proved independently by three different persons in 1935 and 1936) says that

> when we have IID sampling from a statistical model, all distributions in the model have the same support which does not depend on the parameter, and all distributions in the model are continuous, then there is a sufficient statistic whose dimension does not depend on the parameter if and only if the statistical model is an exponential family of distributions.

This theorem was responsible for the interest in exponential families early in the twentieth century. The condition of the Pitman-Koopman-Darmois theorem that the support does not depend on the parameter is essential. For IID sampling from the Uniform$(0, \theta)$ model the maximal order statistic $X_{(n)}$ is sufficient. Its dimension (one) does not depend on n. To show this note that the likelihood is

$$L_n(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{(0,\theta)}(X_i)$$

$$
\begin{aligned}
&= \frac{1}{\theta^n} \prod_{i=1}^{n} I_{(0,\theta)}(X_i) \\
&= \frac{1}{\theta^n} I_{(0,\theta)}(X_{(n)})
\end{aligned}
$$

because if $X_{(n)} < \theta$ then $X_i < \theta$ for all i,

The condition that the statistical model has to be continuous is ugly. Many of the most important applications of exponential family theory (logistic and Poisson regression, log-linear models for categorical data) are discrete, and the theorem does not say anything about them. But later theorems that did cover discrete distributions need extra conditions that seem just there so the theorem can be proved.

Interest in exponential families changed direction in the 1970's with the invention of generalized linear models [Nelder and Wedderburn, 1972, Wedderburn, 1974] and log-linear models for categorical data [Bishop et al., 2007] and with the publication of authoritative treatises [Barndorff-Nielsen, 2014, Brown, 1986] which used the mathematics of convex analysis [Rockafellar, 1970].

In that context the sufficient dimension reduction for canonical linear submodels (exponential family regression models) became more important than the Pitman-Koopman-Darmois property. This is the relation between the canonical sufficient statistic y of the saturated model and the canonical sufficient statistic $M^T y$ of a canonical linear submodel. The former has the row dimension of M and the latter has the column dimension of M, which is usually much smaller.

Maximum Entropy

Entropy is a physical quantity involved in the second law of thermodynamics, which says that the the total entropy of an isolated physical system is nondecreasing in any physical process. It has to do with the maximum possible efficiency of a heat engine or refrigerator, with which chemical reactions proceed spontaneously, and with many other things.

Ludwig Boltzmann and Josiah Willard Gibbs figured out the connection between entropy and probability and between the thermodynamic properties of bulk matter and the motions and interactions of atoms and molecules.

In this theory entropy is not certain to increase to its maximum possible value. It is only overwhelmingly probable to do so in any large system. In a very small system, such as a cubic micrometer of air, it is less probable that entropy will be near its maximum value. In such a small system the statistical fluctuations are large. This is the physical manifestation of the law of large numbers. The larger the sample size (the more molecules involved) the less stochastic variation. Boltzmann thought this discovery so important that he had $S = k \log W$ inscribed on his tombstone (S is entropy, W is probability, and k is a constant of nature now known as Boltzmann's constant).

Claude Shannon imported entropy into information theory, using it to determine the maximum

throughput of a noisy communication channel. Shannon information is negative entropy (minus log probability). Kullback and Leibler imported the same concept into statistics, where it is usually called *Kullback-Leibler information*. It is expected log likelihood and hence what likelihood attempts to estimate.

Edwin Jaynes, a physicist, introduced the "maximum entropy formalism" that describes exponential families in terms of entropy. To keep the derivation simple, we will do the finite sample space case. The same idea can be extended to the infinite discrete case or the continuous case, although the math is harder.

The *relative entropy* of a distribution with PMF f to a distribution with PMF m is defined to be

$$-\sum_{x \in S} f(x) \log\left(\frac{f(x)}{m(x)}\right),$$

where S is the support of the distribution with PMF m. (It is the negative of this quantity that is Kullback-Leibler information of f with respect to m.) It is actually not necessary that m be a PMF; any positive function will do.

Suppose we "know" the value of some expectations

$$\mu_j = \mathrm{E}\left(t_j(X)\right) = \sum_{x \in S} t_j(x) f(x), \qquad j \in J,$$

and we want f to maximize entropy subject to these constraints plus the constraints that f is nonnegative and sums to one. That is, we want to solve the following optimization problem

$$\begin{aligned} \text{maximize } & -\sum_{x \in S} f(x) \log\left(\frac{f(x)}{m(x)}\right) \\ \text{subject to } & \sum_{x \in S} t_j(x) f(x) = \mu_j, \qquad j \in J \\ & \sum_{x \in S} f(x) = 1 \\ & f(x) \geq 0, \qquad x \in S \end{aligned}$$

It turns out that the inequality constraints here are unnecessary. If we solve the problem without requiring f be nonnegative, the solution happens to be nonnegative. But we do need to enforce the equality constraints.

To do that, we use the method of Lagrange multipliers. Multiply each constraint function by a new parameter (Lagrange multiplier) and add to the objective function. This gives the Lagrangian function

$$\mathcal{L}(f) = -\sum_{x \in S} f(x) \log\left(\frac{f(x)}{m(x)}\right) + \sum_{j \in J} \theta_j \sum_{x \in S} t_j(x) f(x) + \psi \sum_{x \in S} f(x)$$

$$= -\sum_{x \in S} f(x) \left[\log\left(\frac{f(x)}{m(x)}\right) - \sum_{j \in J} \theta_j t_j(x) - \psi \right],$$

where θ_j and ψ are the Lagrange multipliers.

Because the domain of f is finite, we can think of it as a vector having components $f(x)$. The Lagrangian is maximized where its first derivative is zero, so we calculate the first partial derivatives as

$$\frac{\partial \mathcal{L}(f)}{\partial f(x)} = -\log\left(\frac{f(x)}{m(x)}\right) + \sum_{j \in J} \theta_j t_j(x) + \psi - 1,$$

setting this equal to zero and solving for $f(x)$ gives

$$f(x) = m(x) \exp\left(\sum_{j \in J} \theta_j t_j(x) + \psi - 1 \right).$$

We then have to find the value of the Lagrange multipliers that make all of the constraints satisfied. In aid of this, define θ to be the vector having components θ_j and $t(x)$ to be the vector having components $t_j(x)$, so that we can write

$$f(x) = m(x) \exp\left(t(x)^T \theta + \psi - 1 \right).$$

In order to satisfy the constraint that the probabilities sum to one we must have

$$e^{\psi - 1} \sum_{x \in S} m(x) e^{t(x)^T \theta} = 1$$

or

$$1 - \psi = \log\left(\sum_{x \in S} m(x) e^{t(x)^T \theta} \right).$$

Now define

$$c(\theta) = \log\left(\sum_{x \in S} m(x) e^{t(x)^T \theta} \right).$$

Then,

$$f(x) = m(x) e^{t(x)^T \theta - c(\theta)},$$

and this is the density of an exponential family! If we think of the Lagrange multipliers θ_j as unknown parameters rather than constants we still have to adjust, then we see that we have an exponential family with canonical statistic vector $t(x)$, canonical parameter vector θ, and cumulant function $c(\theta)$.

Define μ to be the vector with components μ_j. Then we know from exponential family that

$$\mu = \nabla c(\theta) = g(\theta)$$

and g is a one-to-one function (if the exponential family is identifiable, which happens if there are no redundant constraints), so the Lagrange multiplier vector is

$$\theta = g^{-1}(\mu)$$

and, although we do not have a closed form expression for g^{-1}, we can evaluate $g^{-1}(\mu)$ for any μ that is a possible mean-value parameter vector found by optimization. Our use of the maximum entropy argument is a bit peculiar. First we said that we "knew" the expectations

$$\mu = \mathrm{E}\{t(X)\},$$

and wanted to pick out one probability distribution that maximizes entropy and satisfies this constraint. Then we forgot about "knowing" this constraint and said as μ ranges over all possible values we get an exponential family of probability distributions. Also we have to choose a base measure.

Despite this rather odd logic, the maximum entropy argument does say something important about exponential families. Suppose we have a big exponential family (a "saturated model") and are interested in submodels. Examples are Bernoulli regression, Poisson regression, or categorical data analysis. The maximum entropy argument says the canonical linear submodels are the submodels that, *subject to constraining the means of their submodel canonical statistics, leave all other aspects of the data as random as possible*, where "as random as possible" means maximum entropy. Thus these models constrain the means of their canonical statistics and anti-constrain (leave as unconstrained as possible) everything else.

In choosing a particular canonical linear submodel parameterization $\theta = M\beta$ we are, in effect, modeling only the the distribution of the submodel canonical statistic $t(y) = M^T y$, leaving all other aspects of the distribution of y as random as possible given the control over the distribution of $t(y)$.

Overdispersion (before we summarize)

A common explanation for large deviance (or poor fit) is the presence of a few outliers. When large number of points are identified as outliers, they become unexceptional, and it may be the case that the error distribution is misspecified. In the presence of misspecification in the form of overdispersion, the exponential family takes on a different functional form

$$f(y|\theta, \phi) = \exp\left(\frac{\langle y, \theta\rangle - c(\theta)}{a(\phi)} - b(y, \phi)\right), \tag{26}$$

where y and θ are as before, ϕ is a dispersion parameter, and $b(y, \phi)$ is a function of the data y and the dispersion parameter ϕ. From the perspective of the canonical exponential families that we

have motivated throughout, the function $b(y, \phi)$ is similar to the base measure h that was dropped from consideration in log likelihood based arguments that focused on the parameters. Notice that the density (26) is a generalization of the exponential family density which specifies that $a(\phi) = 1$ and $b(y, \phi) = \log(h(y))$. Note that the dispersion parameter can be estimated using

$$\hat{\phi} = \frac{\sum_{i=1}^{n} (y - \hat{\mu}_i)^2 / \hat{\mu}_i}{n - p}.$$

We investigate overdispersion when we cover GLMs for count responses.

Interpretation

So now we can put all of this together to discuss interpretation of regular full exponential families and their canonical linear submodels.

The MLE is unique if it exists (from strict concavity). Existence is a complicated story, and non-existence results in complicated problems of interpretation, which we leave for now. We will revisit non-existence later.

The MLE satisfies the observed equals expected property, either (16) for a saturated model or (21) for a canonical linear submodel.

The sufficient dimension reduction property and maximum entropy property say that $M^T y$ is a sufficient statistic, hence captures all information about the parameter. All other aspects of the distribution of y are left as random as possible; the canonical linear submodel does not constrain them in any way other than its constraints on the expectation of $M^T y$.

A quote from Charlie Geyer:

> "Parameters are meaningless quantities. Only probabilities and expectations are meaningful."

Of course, some parameters are probabilities and expectations, but most exponential family canonical parameters are not. Also note that the same model can be specified by different formulas or different model matrices, so that a particular canonical parameter value does not specify a model uniquely. A quote from Alice in Wonderland (taken from Charlie Geyer):

> 'If there's no meaning in it,' said the King, 'that saves a world of trouble, you know, as we needn't try to find any.'

Realizing that canonical parameters are meaningless quantities "saves a world of trouble." We "needn't try to find any."

Hence our interpretations should be focused on mean value parameters. This conclusion flies in the face the traditional way regression models are taught. In most courses, students are taught

to "interpret" the equation $\theta = M\beta$, or, more precisely, since in lower level courses students aren't assumed to know about matrices, students are taught to interpret this with the matrix multiplication written out explicitly, interpreting equations like

$$\theta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2, \qquad \text{where } i \text{ runs over cases.}$$

The model matrix M determines two linear transformations

$$\beta \mapsto M\beta$$
$$\mu \mapsto M^T\mu$$

We claim, that the second one, which takes saturated model canonical statistic to submodel canonical statistic and saturated model mean value parameter to submodel mean value parameter, is the more important of the two and should lead in interpretation, because the former is about canonical parameters (the meaningless ones) and the latter is about mean value parameters (the meaningful ones). This is especially so in light of the fact that $M^T y = M^T \hat{\mu}$ (observed equals expected) is the only algebraically simple property of maximum likelihood that users can hang an interpretation on. So we need to rethink the way we teach regression and interpret regression when talking to users.

When we do need to think about canonical parameters, the key concept is the multivariate monotone relationship (13) between canonical and mean value parameters. Note that this holds not only for saturated model parameters but also for canonical linear submodel parameters. If, as before, we let $\tau = M^T\mu$ denote the submodel mean value parameter, and τ_1 corresponds to β_1 and τ_2 to β_2, then

$$(\tau_1 - \tau_2)^T(\beta_1 - \beta_2) > 0, \qquad \text{whenever } \tau_1 \neq \tau_2.$$

By standard theory of maximum likelihood, MLEs of all parameters are consistent, efficient (have minimum asymptotic variance), and asymptotically normal, with easily calculated asymptotic variance (inverse Fisher information matrix). Fisher information is easily calculated, (15) is Fisher information for the saturated model canonical parameter θ;

$$\nabla_\beta^2 c(M\beta) = M^T \left[\nabla c(M\beta)\right] M$$

is Fisher information for the submodel canonical parameter β.

The Delta method then gives asymptotic variance matrices for mean value parameters. If $\mu = g(\theta)$, then the asymptotic variance for $\hat{\mu}$ is

$$[\nabla g(\theta)] I(\theta)^{-1} [\nabla g(\theta)]^T = I(\theta) I(\theta)^{-1} I(\theta) = I(\theta)$$

and $M^T I(\theta) M$ is the asymptotic variance for $\hat{\tau}$. These can be used for hypothesis tests and confidence intervals about these other parameters.

Acknowledgments

These notes take materials from Charles Geyer's notes on exponential families and other topics. We also borrow materials from Trevor Park's STAT 426 notes and Agresti [2013].

References

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, 2013.

O. Barndorff-Nielsen. *Information and exponential families: in statistical theory*. John Wiley & Sons, 2014.

Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media, 2007.

L. D. Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

E. L. Lehmann. *Testing Statistical Hypotheses*. New York: John Wiley, 1959.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

R. T. Rockafellar. *Convex analysis*. Number 28. Princeton University Press, 1970.

R. T. Rockafellar and R. J. B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 1998. (The corrected printings contain extensive changes. We used the 3rd corrected printing, 2010.).

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447, 1974.

R. W. M. Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(3):27–32, 1976.