Stat 8053 Lecture Notes
Exponential Families
Charles J. Geyer
September 29, 2014

1 Exponential Families

1.1 Definition

An *exponential family of distributions* is a parametric statistical model having log likelihood

$$l(\theta) = y^T\theta - c(\theta), \tag{1}$$

where y is a vector statistic and θ is a vector parameter. This uses the convention that terms that do not contain the parameter can be dropped from a log likelihood; otherwise such terms might also appear in (1). A statistic y and parameter θ that give a log likelihood of this form are called *canonical* or *natural*. The function c is called the *cumulant function* of the family.

1.2 Non-Uniqueness

The canonical statistic and parameter are not unique.

- Any one-to-one affine function of a canonical statistic is again canonical, although this also changes the canonical parameter and cumulant function.
- Any one-to-one affine function of a canonical parameter is again canonical, although this also changes the canonical statistic and cumulant function.
- Any scalar-valued affine function of the canonical parameter can be added to the cumulant function, although this also changes the canonical statistic.

We usually do not worry about this. We fix one choice of canonical statistic, canonical parameter, and cumulant function and say "the" canonical statistic, "the" canonical parameter, and "the" cumulant function.

1.3 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

$$f_\theta(w) = h(w)e^{Y(w)^T\theta - c(\theta)}$$

and the word "density" here can refer to a probability mass function (PMF) or a probability density function (PDF) in the usual senses of master's level statistics 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 — both arise in practical examples, as we shall see) or to a density with respect to an arbitrary positive measure in the sense of Ph. D. level probability theory. The $h(w)$ arises from any term not containing the parameter that is dropped in going from log densities to log likelihood.

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.

When we look at a ratio of densities, the $h(w)$ cancels, and

$$f_{\theta;\psi}(w) = e^{Y(w)^T(\theta-\psi) - c(\theta) + c(\psi)} \tag{2}$$

is a density of the distribution with parameter value θ with respect to the distribution with parameter value ψ (a Radon-Nikodym derivative for those who have had measure-theoretic probability; everyone else should ignore this comment). For w such that $h(w) = 0$ (2) 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 (2) is defined for such w.

Now, since (2) is everywhere strictly positive, we see that every distribution in the family has the same support.

1.4 Cumulant Functions

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

$$c(\theta) = c(\psi) + \log E_\psi\left\{e^{Y^T(\theta-\psi)}\right\} \tag{3}$$

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 $c(\theta) = \infty$.

1.5 Full Families

Define

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

Then (2) defines 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).

There is literature about so-called "curved exponential families" and other non-full exponential families, but we will not discuss them.

We have just seen that even if the originally given family only had one distribution in it (the one for parameter value ψ), we get the whole full exponential family from it via (3) and (2) and (4). We say that the exponential family is *generated* by any of the distributions in it.

1.6 Moment Generating Functions

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

$$\begin{aligned} m_\theta(t) &= E_\theta\{e^{Y^T t}\} \\ &= E_\psi\{e^{Y^T (t+\theta-\psi) - c(\theta) + c(\psi)}\} \\ &= e^{c(t+\theta) - c(\theta)} \end{aligned} \tag{5}$$

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 (Fristedt and Gray, 1996, Sections 13.5 and 13.6), if the moment generating function exists, then moments of all orders exist and ordinary moments are given by the derivatives of m_θ evaluated at zero. In particular

$$\begin{aligned} E_\theta(Y) &= \nabla m_\theta(0) = \nabla c(\theta) \\ E_\theta(YY^T) &= \nabla^2 m_\theta(0) = \nabla^2 c(\theta) + [\nabla c(\theta)] \cdot [\nabla c(\theta)]^T \end{aligned}$$

1.7 Cumulant Generating Functions

A log moment generating function is called a *cumulant generating function* and its derivatives evaluated at zero are called the *cumulants* of the distribution. Cumulants of order m are polynomial functions of moments of orders up to m and vice versa (Cramér, 1951, Section 15.10).

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

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

where c is the cumulant function. Note that derivatives of the cumulant generating function (6) evaluated at zero are the same as derivatives of the cumulant function c evaluated at θ. Hence the name "cumulant function."

The first and second cumulants of the canonical statistic are

$$\nabla c(\theta) = E_\theta(Y) \tag{7a}$$

$$\begin{aligned} \nabla^2 c(\theta) &= E_\theta(YY^T) - E_\theta(Y) E_\theta(Y)^T \\ &= \operatorname{var}_\theta(Y) \end{aligned} \tag{7b}$$

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

1.8 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. One that is not regular is the Strauss process, a spatial point process (Geyer and Møller, 1994). We won't say anything else about non-regular exponential families.

1.9 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 (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 is concentrated on a hyperplane

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

for some non-zero vector v, then by (3) for any scalar s

$$\begin{aligned} c(\theta + sv) &= c(\psi) + \log E_\psi\bigl\{e^{Y^T(\theta+sv-\psi)}\bigr\} \\ &= c(\psi) + \log E_\psi\bigl\{e^{sa+Y^T(\theta-\psi)}\bigr\} \\ &= sa + c(\theta) \end{aligned}$$

(the second equality being that $Y^T v = a$ with probability one). And plugging into (2) gives

$$f_{\theta+sv;\psi}(w) = e^{Y(w)^T(\theta+sv-\psi)-c(\theta+sv)+c(\psi)} = e^{s(Y^T v-a)} f_{\theta;\psi}(w)$$

hence $f_{\theta+sv;\psi} = f_{\theta;\psi}$ on the support of the family and hence canonical parameters $\theta + sv$ and θ correspond to the same distribution for all θ.

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.*

A direction 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.

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.

I claim 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). Then (3) gives

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

Then (2) gives

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

We simplify this 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

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

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

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

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 $(1, 1, \ldots, 1)$ 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 (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 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.

1.10 Mean Value Parameterization

The mean of the canonical statistic is also a parameter. It is given as a function of the canonical parameter by (7a)

$$\mu = g(\theta) = \nabla c(\theta). \tag{9}$$

Theorem 2. *For a regular 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

$$h(\theta) = \mu^T \theta - c(\theta). \tag{10}$$

The second derivative of this $-\nabla^2 c(\theta)$ is equal to $-\operatorname{var}_\theta(Y)$ by (7b), and this is a negative definite matrix because Y is not concentrated on a hyperplane. Hence (10) is a strictly concave function (Rockafellar and Wets, 1998, Theorem 2.14), and this implies that the maximum of (10) is unique if it exists (Rockafellar and Wets, 1998, Theorem 2.6). Moreover, we know a solution exists because the derivative of (10) is

$$\nabla h(\theta) = \mu - \nabla c(\theta).$$

and we were assuming that $\mu = \nabla c(\theta)$ for some θ.

Moment generating functions are infinitely differentiable at zero. Hence so are cumulant generating functions because the logarithm function is infinitely differentiable. Hence cumulant functions are infinitely differentiable on the interior of the full canonical parameter space. Hence the change-of-parameter (9) is infinitely differentiable.

The Jacobian matrix of the change-of-parameter (9) is

$$\nabla g(\theta) = \nabla^2 c(\theta). \tag{11}$$

If the model is identifiable, then Y is not concentrated on a hyperplane, so its variance matrix is nonsingular, hence by (7b) the Jacobian (11) is nonsingular for all θ.

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 we have already proved exists), and the derivative of the inverse is the inverse of the derivative

$$\nabla g^{-1}(\mu) = \bigl[\nabla g(\theta)\bigr]^{-1}, \qquad \text{when } \mu = g(\theta) \text{ and } \theta = g^{-1}(\mu) \tag{12}$$

Now the formula for the derivative of a matrix inverse

$$\frac{\partial A^{-1}}{\partial t} = -A^{-1}\frac{\partial A}{\partial t}A^{-1},$$

which can be proved by differentiating $AA^{-1} = I$, shows that the matrix inversion is infinitely differentiable, and this shows that g^{-1} is infinitely differentiable. And that proves the theorem.

1.11 Multivariate Monotonicity

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

$$\bigl[g(x_1) - g(x_2)\bigr]^T(x_1 - x_2) \geq 0, \qquad \text{for } x_1 \text{ and } x_2 \text{ in } \mathbb{R}^d, \tag{13}$$

and *strictly multivariate monotone* if (13) holds with strict inequality whenever $x_1 \neq x_2$.

If g is differentiable, then (Rockafellar and Wets, 1998, Proposition 12.3) it is multivariate monotone if and only if the symmetric part of the Jacobian matrix $\nabla g(x)$ 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 $\nabla g(x)$ be positive definite for each x.

If g is the mapping from canonical to mean value parameter (9) 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

$$(\mu_1 - \mu_2)^T(\theta_1 - \theta_2) > 0, \qquad \text{whenever } \mu_1 \neq \mu_2. \tag{14}$$

In general, this is all we can say about the map from canonical to mean value parameters.

There is a dumbed down version of (14). If we rewrite (14) 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 k-th, 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 dumbed down property is not enough to make some arguments about exponential families that are needed in applications (Shaw and Geyer, 2010, Appendix).

Here is another rewrite of (14) that preserves its full force. Fix a vector $v \neq 0$. Write $\theta_2 = \theta$ and $\theta_1 = \theta + sv$, so multivariate monotonicity (13) becomes

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

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*

$$g'(\theta; v) = v^T \big[\nabla g(\theta)\big] v = v^T \big[\nabla^2 c(\theta)\big] v. \tag{15}$$

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 dumbed down property only says that all the partial derivatives are nonzero and this corresponds to asserting (15) 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 dumbed down property is dumbed down. 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.

1.12 Maximum Likelihood

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 are the same

$$I(\theta) = \nabla^2 c(\theta). \tag{16}$$

When the model is identifiable, the second derivative matrix is negative definite everywhere, hence the log likelihood is strictly concave, hence the maximum likelihood estimate is unique if it exists.

1.12.1 Non-Existence of the MLE, Take One

Unlike our proof in Section 1.10, 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.

1.12.2 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

$$y = \nabla c(\theta) = E_\theta(Y).$$

So 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

$$y = \hat{\mu}. \tag{17}$$

1.13 Independent and Identically Distributed

Suppose y_1, ..., 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

$$\begin{aligned} l_n(\theta) &= \sum_{i=1}^{n} \bigl[y_i^T \theta - c(\theta) \bigr] \\ &= \left(\sum_{i=1}^{n} y_i \right)^T \theta - n c(\theta) \end{aligned} \tag{18}$$

and we see that we again have an exponential family with

- canonical statistic $\sum_{i=1}^{n} y_i$,
- cumulant function $\theta \mapsto n c(\theta)$, and
- canonical parameter θ and full canonical parameter space Θ the same as for the originally given family.

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

1.14 Asymptotics of Maximum Likelihood

Rewrite (18) as

$$l_n(\theta) = n \bigl[\bar{y}_n^T \theta - c(\theta) \bigr]$$

so

$$\nabla l_n(\theta) = n \bigl[\bar{y}_n - \nabla c(\theta) \bigr]$$

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

$$\hat{\theta}_n = g^{-1}(\bar{y}_n) \tag{19}$$

where g is the θ to μ mapping given by (9) and g^{-1} is its inverse function, which was proved to exist in Section 1.10. More precisely, (19) holds when

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

By the multivariate central limit theorem (CLT)

$$\sqrt{n}(\bar{y}_n - \mu) \to \text{Normal}\big(0, I(\theta)\big)$$

and we know that g^{-1} is differentiable with the derivative given by (12). So the usual asymptotics of maximum likelihood

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \text{Normal}\big(0, I(\theta)^{-1}\big) \tag{20}$$

is just the multivariate delta method applied to the multivariate CLT. For details see the 8112 notes (Geyer, 2013). In fact, Theorem 10 in those notes asserts a slightly stronger conclusion, that (20) holds for every identifiable full exponential family, whether or not it is regular, so long as the true unknown parameter θ is in the interior of the canonical parameter space.

In summary, one "regularity condition" for (20) to hold is that we have an identifiable regular full exponential family. Of course, (20) holds for many non-exponential-family models, but the regularity conditions are so complicated that they are often hard to verify. In exponential families the verification is trivial: the usual asymptotics of maximum likelihood always works.

1.15 Canonical Affine Submodels

A *canonical affine submodel* of an exponential family is a submodel having parameterization

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

where θ is the canonical parameter of the originally given exponential family, β is the parameter of the submodel, a is a known vector, and M is a known matrix. The matrix M is called the *model matrix* in the terminology used by the R functions `lm` and `glm`. Other people call M the *design matrix*, although this is not really appropriate when data are not from a designed experiment. 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*.

The submodel log likelihood is

$$\begin{aligned} l(\beta) &= y^T(a + M\beta) - c(a + M\beta) \\ &= y^T a + y^T M\beta - c(a + M\beta) \\ &= y^T a + (M^T y)^T \beta - c(a + M\beta) \end{aligned}$$

and the term $y^T a$ can be dropped because it does not contain the parameter β giving log likelihood

$$l(\beta) = (M^T y)^T \beta - c(a + M\beta) \tag{21}$$

and we see that we again have an exponential family with

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

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

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

Thus a canonical affine 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.

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 affine submodel canonical and mean value parameters β and $\tau = M^T \mu$.

The observed equals expected property for the submodel is

$$\hat{\tau} = M^T \hat{\mu} = M^T y. \tag{22}$$

We cannot actually solve these equations for $\hat{\mu}$ because M the mapping $\mu \mapsto M^T \mu$ is usually not one-to-one (the $n > p$ case where M is $n \times p$ and full 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 (21) 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. It is the only simple property of maximum likelihood that can be used in interpretation of exponential families (more on this later, Section 1.19).

1.16 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. It follows that Bayesian inference always obeys the likelihood principle. It also follows that likelihood inference can obey the likelihood principle, although this is not automatic. The maximum likelihood estimator (MLE), the likelihood ratio test statistic, and observed and expected Fisher information with the MLE plugged in all depend on the data only through the likelihood, hence obey the sufficiency principle. Other procedures that are sometimes considered part of likelihood inference, like one-step Newton updates of root-n-consistent estimators, do not necessarily obey the sufficiency principle.

Corollary 3. *The canonical statistic vector of an exponential family is a sufficient statistic.*

As a reminder of this, some statisticians have a habit of saying "canonical sufficient statistic" or "natural sufficient statistic," although this is redundant (the canonical statistic is always sufficient), in order to emphasize the sufficiency property.

1.17 Sufficient Dimension Reduction

Here at Minnesota we hear a lot about *sufficient dimension reduction* (Chiaromonte, Cook, and Li, 2002, and subsequent papers citing this). That is very complicated theory and we shall not discuss it.

But, it is good to remember that 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

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

because if $X_{(n)} < \theta$ then so are $X_i < \theta$ 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 (my brother-in-law's thesis advisor called these "ham and eggs theorems" — if we had some ham, we'd have ham and eggs, if we had some eggs).

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, Fienberg, and Holland, 2007, originally published 1975) and with the publication of authoritative treatises (Barndorff-Nielsen, 1978; Brown, 1986) which used the recently developed mathematics called convex analysis (Rockafellar, 1970).

In that context the sufficient dimension reduction for canonical affine submodels (exponential family regression models) became more important than the Pitman-Koopman-Darmois property. This is (Section 1.15) the relation between the canonical sufficient statistic y of the saturated model and the canonical sufficient statistic $M^T y$ of a canonical affine submodel. The former has the row dimension of M and the latter has the column dimension of M, which is usually much smaller.

1.18 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

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

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

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

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

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

θ_j, $j \in J$, 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 first partial derivatives

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

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

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

Then we have to find the value of the Lagrange multipliers that make all 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 we can write

$$f(x) = m(x) e^{t(x)^T \theta + \psi - 1}$$

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

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

That looks familiar!

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.

Define μ to be the vector with components μ_j. Then we know from exponential family theory 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 of the mean value parameter vector by doing an optimization.

Our use of the maximum entropy argument is a bit peculiar. First we said that we "knew" the expectations

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

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 affine 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 constraint the means of their canonical statistics and anti-constrain (leave as unconstrained as possible) everything else.

In choosing a particular canonical affine submodel parameterization $\theta = a + M\beta$ 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 $t(y)$.

1.19 Interpretation

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

The MLE satisfies the observed equals expected property, either (17) for a saturated model or (22) for a canonical affine 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 affine submodel does not constrain them in any way other than its constraints on the expectation of $M^T y$.

A quote from my master's level theory notes

> 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.

A quote from *Alice in Wonderland*

> '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 $\theta = a + M\beta$, 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

$$\begin{aligned} \beta &\mapsto M\beta \\ \mu &\mapsto M^T\mu \end{aligned}$$

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 (14) between canonical and mean value parameters. Note that this holds not only for saturated model parameters but also for canonical affine 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

$$(\tau_1 - \tau_2)^T(\beta_1 - \beta_2) > 0, \qquad \text{whenever } \tau_1 \neq \tau_2. \tag{23}$$

By standard theory of maximum likelihood, MLE's of all the 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, (16) is Fisher information for the saturated model canonical parameter θ;

$$\nabla_\beta^2 c(a + M\beta) = M^T \bigl[\nabla^2 c(a + M\beta)\bigr] M$$

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 $\hat{\mu}$

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

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.

References

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families*. Wiley, Chichester.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete Multivariate Analysis: Theory and Applications*. Springer, New York. Originally published by MIT Press, 1975.

Browder, A. (1996). *Mathematical Analysis: An Introduction*. Springer-Verlag, New York.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA.

Chiaromonte, F., Cook, R. D., and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Annals of Statistics*, **30**, 475–497.

Cramér, H. (1951). *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

Fristedt, B. E. and Gray, L. F. (1996). *A Modern Approach to Probability Theory*. Boston: Birkhäuser.

Geyer, C. J. (1990). Likelihood and Exponential Families. PhD thesis, University of Washington. `http://purl.umn.edu/56330`.

Geyer, Charles J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, **3**, 259–289.

Geyer, Charles J. (2013). Stat 8112 Lecture Notes: Asymptotics of Exponential Families. `http://www.stat.umn.edu/geyer/8112/notes/expfam.pdf`.

Geyer, C. J. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359–373.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. John Wiley & Sons, New York.

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

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.

Rockafellar, R. T., and Wets, R. J.-B. (1998). *Variational Analysis*, Springer-Verlag, Berlin. (The corrected printings contain extensive changes. We used the 3rd corrected printing, 2010.)

Shaw, R. G., and Geyer, C. J. (2010). Inferring fitness landscapes *Evolution*, **64**, 2510–2520.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.