Stat 8112 Lecture Notes

Asymptotics of Exponential Families

Charles J. Geyer

January 23, 2013

1 Exponential Families

An *exponential family of distributions* is a parametric statistical model having densities with respect to some positive measure λ of the form

$$f_\alpha(\omega) = a(\omega) \exp\left(\sum_{i=1}^{d} Y_i(\omega)\theta_i(\alpha) - b(\alpha) \right) \tag{1}$$

where a, b, Y_i, and θ_i are real-valued functions and a must be nonnegative. If the densities of a statistical model can be put in the form (1), then there are many ways to do so. Each is called a *representation* of the family. A representation is *minimal* if d is as small as possible.

It simplifies notation if we consider $Y = (Y_1, \ldots, Y_d)$ and $\theta = (\theta_1, \ldots, \theta_d)$ as vectors, and write

$$\langle y, \theta \rangle = \sum_{i=1}^{d} y_i \theta_i.$$

If we define

$$c(\theta) = \log \int a(\omega) e^{\langle Y(\omega), \theta \rangle} \, \lambda(d\omega). \tag{2}$$

then we can rewrite the densities (1) as

$$f_\theta(\omega) = a(\omega) e^{\langle Y(\omega), \theta \rangle - c(\theta)}. \tag{3}$$

The log likelihood is

$$l(\theta) = \langle y, \theta \rangle - c(\theta) \tag{4}$$

if we use the convention that additive terms that do not contain the parameter may be dropped from log likelihoods (because they do not affect any likelihood-based statistical inferences). Thus we have another definition of exponential families: a statistical model is an *exponential family of distributions* if there exists a statistic Y and a parameter θ such that the log likelihood has the form (4).

The statistic and parameter that put the log likelihood in the exponential family form (4) or the densities in the exponential family form (3) are called the *natural statistic* and *natural parameter* (alternative terminology is *canonical statistic* and *canonical parameter*). The function c is called the *cumulant function* of the family.

The probability measures in the family are given by

$$P_\theta(B) = \int_B f_\theta \, d\lambda, \qquad B \in \mathcal{B},$$

where $\mathcal{B}$ is the sigma-algebra that is the domain of λ.

2 Minimality

A *hyperplane* in $\mathbb{R}^d$ is a set

$$H = \{\, y \in \mathbb{R}^d : \langle y, \varphi \rangle = b \,\}$$

for some nonzero vector φ and some real number b.

Theorem 1. *Every exponential family has a minimal representation, and a representation is minimal if and only if neither of the following conditions hold.*

(i) *There exists a hyperplane that contains the natural statistic vector with probability one.*

(ii) *There exists a hyperplane that contains the natural parameter space.*

It will take some set-up before we can prove the theorem.

We may choose the measure λ to be one of the probability measures in the family, say P_ψ. The density of P_θ with respect to P_ψ is given by the ratio of the corresponding densities (3), that is,

$$g_\theta(\omega) = e^{\langle Y(\omega), \theta - \psi \rangle - c(\theta) + c(\psi)} \tag{5}$$

(it does no harm to take $a(\omega)/a(\omega) = 1$ even when $a(\omega) = 0$ because the set of such ω has P_ψ measure zero, and we may redefine densities on sets of measure zero without changing the distribution).

The fact that the densities (5) of one distribution in the family with respect to another are strictly positive means that all distributions in the family have the same null sets: $P_\theta(A) = 0$ if and only if $P_\psi(A) = 0$. Thus

when we say "almost surely" it does not matter which distribution in the family we refer to.

The densities (5) have logarithms

$$h_\theta(\omega) = \langle Y(\omega), \theta - \psi \rangle - c(\theta) + c(\psi) \tag{6}$$

Let $\mathcal{H}$ denote the set of measurable real-valued functions on the sample space. For any $h \in \mathcal{H}$, define

$$[h] = \{\, k \in \mathcal{H} : h - k \text{ is constant almost surely} \,\}.$$

When h_θ is given by (6), we say that $[h_\theta]$ is an *equivalence class of log unnormalized densities* and say that h is a *representative* of the equivalence class $[h]$.

There is a one-to-one correspondence between between probability distributions in the model and equivalence classes of log unnormalized densities. Each representative h of an equivalence class $[h_\theta]$ corresponds to the distribution having density $e^h / E_\psi(e^h)$. Conversely, all densities with respect to P_ψ of one distribution have logarithms in the same equivalence class.

We can consider equivalence classes of log unnormalized densities as vectors by defining vector addition and scalar multiplication in the obvious way

$$[h] + [k] = [h + k]$$
$$r \cdot [h] = [rh]$$

where h and k are elements of $\mathcal{H}$ and r is a scalar.

Proof of Theorem 1. To say that a statistical model is an exponential family is the same thing as saying that the vector space V spanned by the equivalence classes of log unnormalized densities of distributions in the exponential family is also spanned by the equivalence classes $[Y_1]$, ..., $[Y_d]$ containing the components of the natural statistic. Clearly, the representation is minimal when the dimension of V is equal to d, so the vectors $[Y_1]$, ..., $[Y_d]$ form a basis for V.

Saying these vectors are linearly independent is the same as saying there do not exist scalars t_1, ..., t_d not all zero such that

$$\sum_{i=1}^{d} t_i [Y_i] = [0],$$

and this is the same as saying there do not exist scalars t_1, ..., t_d not all zero and another scalar k such that

$$\sum_{i=1}^{d} t_i Y_i = k, \qquad \text{almost surely,} \tag{7}$$

and (7) is the same thing as condition (i) of the theorem. Hence if the representation is minimal, then condition (i) does not hold.

Saying that the dimension of V is d the same as saying there are equivalence classes $[h_{\theta_1}]$, ..., $[h_{\theta_d}]$ where θ_1, ..., θ_d are points in the natural parameter space and h_θ is given by (6) and these equivalence classes are linearly independent, and this is the same as saying there do not exist scalars t_1, ..., t_d not all zero such that

$$\sum_{i=1}^{d} t_i [h_{\theta_i}] = [0].$$

Now

$$\begin{aligned} \sum_{i=1}^{d} t_i [h_{\theta_i}] &= \left[\sum_{i=1}^{d} t_i h_{\theta_i} \right] \\ &= \left[\sum_{i=1}^{d} t_i \langle Y, \theta_i - \psi \rangle \right] \\ &= \left[\left\langle Y, \sum_{i=1}^{d} t_i (\theta_i - \psi) \right\rangle \right] \end{aligned}$$

and to say this is not equal to [0] is to say that

$$\left\langle Y, \sum_{i=1}^{d} t_i (\theta_i - \psi) \right\rangle$$

is not constant almost surely. We have already shown that if the representation is minimal $\langle Y, s \rangle$ being constant almost surely implies $s = 0$. Hence the representation being minimal implies

$$\sum_{i=1}^{d} t_i \theta_i = \sum_{i=1}^{d} t_i \psi, \tag{8}$$

and (8) is the same thing as condition (ii) of the theorem. Hence if the representation is minimal, then condition (ii) does not hold.

Conversely, if condition (i) holds, then (7) holds with some t_i not equal to zero, say $t_j \neq 0$. Then

$$Y_j = \frac{1}{t_j} \left[k - \sum_{i \neq j} t_i Y_i \right], \qquad \text{almost surely,}$$

and (4) can be rewritten as

$$\begin{aligned} l(\theta) &= \frac{1}{t_j} \left[k - \sum_{i \neq j} t_i Y_i \right] \theta_j + \sum_{i \neq j} Y_i \theta_i - c(\theta) \\ &= \sum_{i \neq j} Y_i \left[\theta_i - \frac{t_i}{t_j} \theta_j \right] - c(\theta) + \frac{k}{t_j} \theta_j \end{aligned}$$

which has the exponential family form (4) with different natural parameter vector and cumulant function. Thus condition (i) implies the representation cannot be minimal.

For the other part of the proof of the converse, if condition (ii) holds, then (8) holds with some t_i not equal to zero, say $t_j \neq 0$, and let k denote the right-hand side of (8). Then

$$\theta_j = \frac{1}{t_j} \left[k - \sum_{i \neq j} t_i \theta_i \right],$$

and (4) can be rewritten as

$$\begin{aligned} l(\theta) &= \frac{1}{t_j} \left[k - \sum_{i \neq j} t_i \theta_i \right] Y_j + \sum_{i \neq j} Y_i \theta_i - c(\theta) \\ &= \sum_{i \neq j} \left[Y_i - \frac{t_i}{t_j} Y_j \right] \theta_i - c(\theta) + \frac{k}{t_j} Y_j \end{aligned}$$

which has the exponential family form (4) with different natural statistic vector if we drop the term kY_j/t_j that does not contain the parameter (which we are allowed to do with log likelihoods). Thus condition (ii) implies the representation cannot be minimal. □

3 Convex Functions

The *extended real number* system consists of the real numbers and two additional "numbers" $-\infty$ and $+\infty$. As sets, the real numbers are denoted $\mathbb{R}$ and the extended real numbers are denoted $\overline{\mathbb{R}}$.

It turns out to be very useful to define convex functions taking values in $\overline{\mathbb{R}}$. These are called extended-real-valued convex functions. Of course we are mostly interested in their behavior where they are real-valued, but allowing the values $+\infty$ and $-\infty$ turns out to be a great convenience.

For any function $f : S \to \overline{\mathbb{R}}$, where S is any set,

$$\operatorname{dom} f = \{\, x \in S : f(x) < +\infty \,\}$$

is called the *effective domain* of f. Such a function is said to be *proper* if its effective domain is nonempty and it is real-valued on its effective domain, or, what is equivalent, f is proper if $-\infty < f(x)$ for all x and $f(x) < +\infty$ for at least one x.

Note that these two definitions (of "effective domain" and "proper") treat plus and minus infinity very differently. The reason is that the theory of convex functions finds most of its applications in minimization problems (that is, the object is to minimize a convex function) and minimization too treats plus and minus infinity very differently.

A subset S of a vector space is *convex* if

$$sx + (1 - s)y \in S, \qquad x, y \in S \text{ and } 0 < s < 1.$$

Note that it would make no difference if the definition were changed by replacing $0 < s < 1$ with $0 \le s \le 1$.

An extended-real-valued function f on a vector space is *convex* if

$$f\big(sx + (1 - s)y\big) \le sf(x) + (1 - s)f(y), \qquad x, y \in \operatorname{dom} f \text{ and } 0 < s < 1$$

The inequality in this formula is also known as the *convexity inequality*. Note that on the right-hand side of the convexity inequality neither term can be $+\infty$ because of the requirement $x, y \in \operatorname{dom} f$. Thus there is no need to define what we mean by $\infty - \infty$ in order to use this definition. Similarly since we are requiring $0 < s < 1$ there is no need to define what we mean by $0 \cdot \infty$. All we need are the obvious rules of arithmetic $s \cdot (-\infty) = -\infty$ when $0 < s < \infty$ and $-\infty + x = -\infty$ when $x < +\infty$. Together they imply that the right-hand side of the convexity inequality is $-\infty$ whenever either $f(x)$ or $f(y)$ is $-\infty$.

4 Concave Functions

An extended-real-valued function f is said to be *concave* if $-f$ is convex. For concave functions we change the definitions of "effective domain" and "proper." Instead of applying the original definitions to f, we say that the effective domain of a concave function f is the same as the effective domain of the convex function $-f$ and, similarly, that a concave function f is proper if and only if the convex function $-f$ is proper. The reason is that the theory of concave functions finds most of its applications in maximization problems (that is, the object is to maximize a concave function).

In fact, we only need one theory. If interested in maximizing a concave function, stand on your head and you are minimizing a convex function. The difference in the two situations is entirely trivial, a mere change in terminology and notation. Nothing of mathematical significance changes, which is why we want our definitions of "effective domain" and "proper" to match. Why take convex functions as the main notion and treat concave functions as the secondary notion that cannot be properly understood without reference to the other? Tradition and the way most optimization books are written.

5 Cumulant Functions

In order that the densities (5) integrate to one

$$c(\theta) = c(\psi) + \log E_\psi\bigl\{e^{\langle Y, \theta - \psi\rangle}\bigr\} \tag{9}$$

must hold. We can use (9) to define the cumulant function as an extended-real-valued function defined on all of $\mathbb{R}^d$. At points where the expectation in (9) does not exist in the conventional sense, we define $c(\theta) = \infty$. This makes sense because $\log x \to \infty$ as $x \to \infty$. Since the integrand in (9) is strictly positive, the integral is also strictly positive.[1] Thus we never have

[1] Define

$$A_\epsilon = \{\, \omega \in \Omega : e^{\langle Y(\omega), \theta - \psi\rangle} \geq \epsilon \,\}$$

then

$$A_\epsilon \uparrow \Omega, \qquad \text{as } \epsilon \downarrow 0$$

hence by continuity of probability

$$P_\psi(A_\epsilon) \uparrow 1, \qquad \text{as } \epsilon \downarrow 0$$

Hence there exists an $\epsilon > 0$ such that $P_\psi(A_\epsilon)$ is strictly positive. Now observe

$$E_\psi\bigl\{e^{\langle Y, \theta - \psi\rangle}\bigr\} \geq E_\psi\bigl\{\epsilon I_{A_\epsilon}\bigr\} = \epsilon P_\psi(A_\epsilon)$$

$c(\theta) = -\infty$.

A function f on a metric space is *lower semicontinuous* (LSC) at x if

$$\liminf_{n\to\infty} f(x_n) \geq f(x), \qquad \text{for all sequences } x_n \to x.$$

A function f is LSC if it is LSC at all points of its domain. A function f on a metric space is *upper semicontinuous* (USC) at x if

$$\limsup_{n\to\infty} f(x_n) \leq f(x), \qquad \text{for all sequences } x_n \to x.$$

A function f is USC if it is USC at all points of its domain.

Theorem 2. *The cumulant function of an exponential family is a lower semicontinuous proper convex function.*

Proof. That $c(\theta)$ is never $-\infty$ was established in the preceding discussion (footnote 1). Since $c(\theta)$ is finite for all θ in the natural parameter space (which is nonempty), c is proper. LSC follows from Fatou's lemma, and convexity from Hölder's inequality. □

The effective domain of the cumulant function is

$$\Theta = \{\, \theta \in \mathbb{R}^d : c(\theta) < \infty \,\}. \tag{10}$$

The actual natural parameter space must be a subset of Θ. Conversely, for every $\theta \in \Theta$, there exists a distribution having density (5) with respect to P_ψ, and the collection of all such distributions is an exponential family with natural statistic y, natural parameter θ, cumulant function c, and natural parameter space Θ. This largest possible family having this natural statistic, natural parameter, and cumulant function is called the *full* family, that is, an exponential family is *full* if its natural parameter space is (10).

Corollary 3. *The natural parameter space of a full exponential family is a convex set.*

Corollary 4. *The log likelihood for the natural parameter of an exponential family is an upper semicontinuous concave function.*

---

which is strictly positive.

A full exponential family is *regular* if its natural parameter space (10) is an open set. Most exponential families that occur in applications are regular.[2]

An exponential family is *convex* (also called *flat*) if its natural parameter space is a convex subset of the full natural parameter space ($\operatorname{dom} c$, where c is the cumulant function). It is *closed convex* if its log likelihood is an upper semicontinuous convex function, which happens when its natural parameter space is the intersection of a closed convex set and the full natural parameter space ($\operatorname{dom} c$).[3]

An exponential family is *curved* if it is a smooth submodel of a full exponential family that is not itself a flat exponential family, where *smooth* means the natural parameter space is specified as the image of a twice continuously differentiable function from $\mathbb{R}^p$ for some p into the full natural parameter space.[4]

6 Identifiability

A parametric family of probability distributions is *identifiable* if there do not exist distinct parameter values corresponding to the same distribution.

Theorem 5. *A full exponential family is identifiable if and only if condition (i) of Theorem 1 does not hold. Moreover, if θ and ψ are distinct parameter values corresponding to the same distribution then $s\theta + (1 - s)\psi$ is contained in the full natural parameter space and corresponds to the same distribution for all real s.*

So, if one uses a minimal representation, then the family is identifiable.

Proof. The distributions P_θ and P_ψ are the same if and only if the density (5) of P_θ with respect to P_ψ is equal almost surely to the density of P_ψ with

[2] A rare example of a non-regular exponential family that occurs in an actual application is the Strauss process, a spatial point process. Geyer and Møller (1994) show that the natural parameter space for this exponential family is $\{\, \theta \in \mathbb{R}^2 : \theta_2 \le 0 \,\}$ (their Example 2 of Section 3) and also show that this necessitates using constrained maximum likelihood (fifth paragraph of their Section 5) so this model does not follow the theory presented in this handout: constrained maximum likelihood is necessary to deal with the possibility that the MLE is on the boundary of the natural parameter space.

[3] Closed convex exponential families require constrained maximum likelihood.

[4] Curved exponential families behave better in some respects than arbitrary statistical models, but do not behave as well as full exponential families or even as well as closed convex exponential families. Thus the theory of curved exponential families is also deferred to another handout.

respect to itself, which is equal to one. That is,

$$e^{\langle Y,\theta-\psi\rangle-c(\theta)+c(\psi)} = 1, \qquad \text{almost surely,}$$

or, because the log function is one-to-one,

$$\langle Y,\theta-\psi\rangle = c(\theta)-c(\psi), \qquad \text{almost surely,} \tag{11}$$

and this says Y is concentrated on a hyperplane.

The "moreover" also follows from (5). Suppose (11) holds. Then, for any bounded random variable X,

$$\begin{aligned} E_{s\theta+(1-s)\psi}(X) &= E_\psi\{Xe^{\langle Y,s\theta+(1-s)\psi-\psi\rangle-c\left(s\theta+(1-s)\psi\right)+c(\psi)}\} \\ &= e^{-c\left(s\theta+(1-s)\psi\right)+c(\psi)}E_\psi\{Xe^{s\langle Y,\theta-\psi\rangle}\} \\ &= e^{-c\left(s\theta+(1-s)\psi\right)+c(\psi)+s[c(\theta)-c(\psi)]}E_\psi(X) \end{aligned}$$

Taking X to be the constant random variable everywhere equal to one, we get

$$c\big(s\theta+(1-s)\psi\big) = sc(\theta)+(1-s)c(\psi), \qquad s\in\mathbb{R} \tag{12}$$

and plugging this back into the above gives

$$E_{s\theta+(1-s)\psi}(X) = E_\psi(X)$$

for all bounded random variables X, which says $s\theta+(1-s)\psi$ and ψ correspond to the same probability distribution. Equation (12) also implies $s\theta+(1-s)\psi\in\operatorname{dom}c$ for all real s. □

7 Affine Change of Parameter

A function $g:\mathbb{R}^p\to\mathbb{R}^d$ is *affine* if it has the form

$$g(\varphi) = M\varphi+a$$

for some linear transformation $M:\mathbb{R}^p\to\mathbb{R}^d$ and some $a\in\mathbb{R}^d$. We now want to consider such an affine change of natural parameter for an exponential family. Writing $\theta=g(\varphi)$ we have

$$\langle y,\theta\rangle = \langle y,M\varphi\rangle+\langle y,a\rangle = \langle M^Ty,\varphi\rangle+\langle y,a\rangle$$

where $M^T : \mathbb{R}^d \to \mathbb{R}^p$ is the *adjoint* of M, which the linear operator represented by the the transpose of the matrix that represents M. This is obvious from

$$\langle y, A\varphi \rangle = y^T A\varphi = (A^T y)^T \varphi = \langle A^T y, \varphi \rangle \tag{13}$$

switching back and forth between the notation of bilinear forms and linear transformations and matrix notation. There is a abuse of notation in the right-hand side of (13) in that the bilinear forms $\langle M^T y, \varphi \rangle$ and $\langle y, a \rangle$ have different dimensions, p and d, respectively, but the meaning is clear.

The log likelihood in the new parameterization is

$$l(\varphi) = \langle M^T y, \varphi \rangle - c(M\varphi + a)$$

(we have dropped the term $\langle y, a \rangle$, which does not contain the parameter). Thus we have a new exponential family with natural statistic $M^T Y$, natural parameter φ, and cumulant function $\varphi \mapsto c(M\varphi + a)$, where c is the original cumulant function.

In short, an affine change of natural parameter gives another exponential family. This is the reason why natural affine submodels (also called canonical affine submodels) are important in exponential family theory. In the terminology of the R function `glm` the matrix M is called the *model matrix* and the vector a is called the *offset vector*. Others call the model matrix the "design matrix" whether or not the model arises from a designed experiment. We will use "model matrix."

The terminology "generalized linear model" should really be "generalized affine model" when the offset vector is not zero, but nobody says this. Most applications have $a = 0$, so "generalized linear model" is appropriate most of the time.

8 Independent, Identically Distributed Data

Another important application that does not take us out of exponential families is independent and identically distributed data. If Y_1, Y_2, ... are IID exponential family natural statistics, then the log likelihood for sample size n is

$$l_n(\theta) = \left\langle \sum_{i=1}^n y_i, \theta \right\rangle - nc(\theta)$$

Thus we have a new exponential family with natural statistic $\sum_{i=1}^n y_i$, natural parameter θ, and cumulant function $\theta \mapsto nc(\theta)$, where c is the original cumulant function.

9 Moment Generating Functions

The *moment generating function* of the natural statistic, if it exists, is given by (5)

$$\begin{aligned} m_\theta(t) &= E_\theta\{e^{\langle Y, t\rangle}\} \\ &= E_\psi\left\{e^{\langle Y, t+\theta-\psi\rangle - c(\theta)+c(\psi)}\right\} \\ &= e^{c(t+\theta)-c(\theta)} \end{aligned}$$

The moment generating function exists if it is finite on a neighborhood of zero, that is, if $\theta \in \operatorname{int}(\operatorname{dom} c)$. If $\theta \notin \operatorname{int}(\operatorname{dom} c)$, then m_θ is not a moment generating function.

By the theory of moment generating functions (Fristedt and Gray, 1996, Sections 13.5 and 13.6), if $\theta \in \operatorname{int}(\operatorname{dom} c)$, then moments of the natural statistic of all orders exist and ordinary moments are given by the derivatives of m_θ evaluated at zero. In particular

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

A log moment generating function is called a *cumulant generating function* and its derivatives evaluated at zero are called the *cumulants* of the distribution. If $\theta \in \operatorname{int}(\operatorname{dom} c)$, then the cumulant generating function of the natural statistic is

$$t \mapsto c(t+\theta) - c(\theta), \tag{14}$$

where c is the cumulant function. Note that derivatives of the cumulant generating function (14) evaluated at zero are the same as derivatives of the cumulant function c evaluated at θ. Hence the name "cumulant function." Cumulants of order m are polynomial functions of moments of orders up to m and vice versa (Cramér, 1951, Section 15.10). For an exponential family, the first and second cumulants of the natural statistic are

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

In short, the mean and variance of the natural statistic always exist when $\theta \in \operatorname{int}(\operatorname{dom} c)$ and are given by derivatives of the cumulant function.

For a regular full exponential family $\operatorname{int}(\operatorname{dom} c)$ is the natural parameter space. As we have said, most applications involve regular full exponential families. In them every distribution has moments of all orders for the natural statistic given by derivatives of the cumulant function.

10 Maximum Likelihood

A point x is a *global minimizer* of an extended-real-valued function f on $\mathbb{R}^d$ if

$$f(y) \geq f(x), \qquad y \in \mathbb{R}^d,$$

and x is a *local minimizer* if there exists a neighborhood U of x such that

$$f(y) \geq f(x), \qquad y \in U.$$

Lemma 6. *For an extended-real-valued proper convex function, every local minimizer is a global minimizer.*

The proof is left as a homework problem.

Corollary 7. *For an extended-real-valued proper concave function, every local maximizer is a global maximizer.*

Lemma 8. *In an identifiable convex exponential family, the maximum likelihood estimate is unique if it exists.*

Proof. If θ_1 and θ_2 are distinct maximizers of the log likelihood, then concavity of the log likelihood implies

$$l(s\theta_1 + (1-s)\theta_2) \geq sl(\theta_1) + (1-s)l(\theta_2), \qquad 0 \leq s \leq 1. \tag{15}$$

But since the value at a global maximizer cannot be exceeded, we must have equality in (15), that is,

$$\langle y, s\theta_1 + (1-s)\theta_2 \rangle - c(s\theta_1 + (1-s)\theta_2) = l(\theta_1), \qquad 0 \leq s \leq 1.$$

Hence

$$\frac{d^2}{ds^s} c(s\theta_1 + (1-s)\theta_2) = 0, \qquad 0 < s < 1.$$

The map $s \mapsto s\theta_1 + (1-s)\theta_2$ is an affine change of parameter, hence by the theory of Section 7 it induces a one-parameter exponential family with natural parameter s, natural statistic $\langle Y, \theta_1 - \theta_2 \rangle$, and cumulant function $s \mapsto c(s\theta_1 + (1-s)\theta_2)$. By the theory of moment generating functions, this cumulant function having second derivative zero means that the natural statistic $\langle Y, \theta_1 - \theta_2 \rangle$ has variance zero, which means that Y itself is concentrated on a hyperplane and the original family is not identifiable, contrary to hypothesis. Thus we have reached a contradiction and hence the assumption that distinct maximizers exist is false. □

If the log likelihood of an identifiable full exponential family is maximized in $\operatorname{int}(\operatorname{dom} c)$ then its gradient is zero there. That is, the MLE is the unique solution (unique by Lemma 8) of

$$\nabla l(\theta) = y - \nabla c(\theta) = 0,$$

that is, the θ such that

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

A function $f : U \to V$ where U and V are open sets in $\mathbb{R}^d$ is a *diffeomorphism* if it is a continuously differentiable function and has a continuously differentiable inverse.

Lemma 9. *If c is the cumulant function of an identifiable full exponential family then the function $\tau : \operatorname{int}(\operatorname{dom} c) \to \mathbb{R}^d$ defined by*

$$\tau(\theta) = \nabla c(\theta) = E_\theta(Y)$$

is a diffeomorphism onto its range, which is an open set, and

$$\nabla \tau(\theta) = \nabla^2 c(\theta) = \operatorname{var}_\theta(Y) \tag{16a}$$

and

$$\nabla \tau^{-1}(\mu) = \left(\nabla \tau(\theta)\right)^{-1}, \qquad \text{when } \mu = \tau(\theta). \tag{16b}$$

Proof. An argument very similar to the proof of Lemma 8 shows that τ is one-to-one. Suppose $\mu = \tau(\theta_1) = \tau(\theta_2)$. Then both θ_1 and θ_2 are maximizers of the function λ defined by

$$\lambda(\theta) = \langle \mu, \theta \rangle - c(\theta)$$

(which is the same as the log likelihood except that μ is not necessarily a possible value of the natural statistic, just a possible value of the *expectation* of the natural statistic). Then the same argument as the proof of Lemma 8 with l replaced by λ shows that $\theta_1 = \theta_2$. This shows τ is an invertible function from $\operatorname{int}(\operatorname{dom} c)$ to its range. Also (16a) holds by the theory of moment generating functions, so τ is differentiable, and the derivative is nonsingular because Y is not concentrated on a hyperplane. Thus by the inverse function theorem (Browder, 1996, Theorems 8.15 and 8.27) if $\mu = \tau(\theta)$ then τ has an inverse defined on some neighborhood of μ such that (16b) holds (we actually know that τ has an inverse defined on the entire range of τ). Since μ was any point of the range of τ, this implies the range of τ is open. □

Let W denote the range of τ. We would now like to define the MLE to be $\tau^{-1}(y)$, and that is fine for $y \in W$, but the MLE is undefined when $y \notin W$. Thus we extend the domain of definition by adding a point u (for "undefined") to the set of possible MLE values, which we consider an isolated point of this set,[5] and define

$$\hat{\theta}(y) = \begin{cases} \tau^{-1}(y), & y \in W \\ u, & \text{otherwise} \end{cases}$$

Note that this function is measurable, because it is continuous on the open set $W = \tau(\operatorname{int}(\operatorname{dom} c))$ and constant (equal to u) on the complement of W.

Now we consider IID data Y_1, Y_2, ... As we saw in Section 8, the log likelihood is

$$l_n(\theta) = n\langle \bar{y}_n, \theta \rangle - nc(\theta)$$

and the MLE is $\hat{\theta}_n = \hat{\theta}(\bar{y}_n)$.

Theorem 10. *For a full exponential family with cumulant function c, if the true parameter value θ_0 is in* $\operatorname{int}(\operatorname{dom} c)$ *and the family is identifiable, then the maximum likelihood estimate $\hat{\theta}_n$ exists with probability converging to one, is unique when it exists, and*

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{w} \text{Normal}\big(0, I(\theta_0)^{-1}\big)$$

where $I(\theta) = \nabla^2 c(\theta)$.

Proof. Since θ_0 is in $\operatorname{int}(\operatorname{dom} c)$, moments of the natural statistic of all orders exist, in particular,

$$\begin{aligned} \mu_0 &= E_{\theta_0}(Y) = \tau(\theta_0) = \nabla c(\theta_0) \\ I(\theta_0) &= \operatorname{var}_{\theta_0}(Y) = \nabla \tau(\theta_0) = \nabla^2 c(\theta_0) \end{aligned}$$

where τ is the function defined above. Then the function $\hat{\theta}$ defined above is differentiable at μ_0 by Lemma 9 and has derivative

$$\nabla \hat{\theta}(\mu_0) = \nabla \tau^{-1}(\mu_0) = I(\theta_0)^{-1}.$$

[5] One way to do this and still remain in a Euclidean space is to embed W in $\mathbb{R}^{d+1}$ as the set

$$\widetilde{W} = \{\, (y_1, \ldots, y_d, 0) : (y_1, \ldots, y_d) \in W \,\}$$

and defining

$$u = (0, \ldots, 0, 1).$$

By the multivariate central limit theorem

$$\sqrt{n}\left(\overline{Y}_n - \mu_0\right) \xrightarrow{w} \text{Normal}\left(0, I(\theta_0)\right)$$

Thus the multivariate delta method implies

$$\sqrt{n}\left[\hat{\theta}(\overline{Y}_n) - \hat{\theta}(\mu_0)\right] \xrightarrow{w} I(\theta_0)^{-1} Z$$

where $Z \sim \mathcal{N}\left(0, I(\theta_0)\right)$. Since

$$\text{var}\{I(\theta_0)^{-1} Z\} = I(\theta_0)^{-1} I(\theta_0) I(\theta_0)^{-1} = I(\theta_0)^{-1}$$

we are done, the assertion about the probability of the MLE existing with probability converging to one following from the portmanteau theorem and the assertion about the uniqueness of the MLE following from Lemma 8. □

References

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

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