18.466 Notes, Feb. 26, 2013, R. Dudley

Exponential families

These will be families $\{P_\theta\}_{\theta\in\Theta}$ of laws, including many of the best-known special families such as the binomial and normal laws, and for which there is a natural vector-valued sufficient statistic, whose dimension stays constant as the sample size n increases. In these notes "law" always means a probability measure.

Definition. A family $\mathcal{P} = \{Q_\psi : \psi \in \Psi\}$ of laws on a measurable space $(X, \mathcal{B})$, containing at least two different laws, is called an *exponential family* if there exist a σ-finite measure μ on $(X, \mathcal{B})$, a positive integer k, and real functions θ_j on Ψ and measurable h with $0 < h(x) < \infty$ and T_j on X for $j = 1, \ldots, k$, such that for all $\psi \in \Psi$, Q_ψ is absolutely continuous with respect to μ, and for some $C(\theta(\psi)) > 0$, where $\theta(\psi) := (\theta_1(\psi), \ldots, \theta_k(\psi))$, $(dQ_\psi/d\mu)(x) = f(\psi, x)$ where

$$(1)\quad f(\psi, x) := (dQ_\psi/d\mu)(x) = C(\theta(\psi))h(x)\exp\left(\sum_{j=1}^{k}\theta_j(\psi)T_j(x)\right).$$

In other words $f(\psi, x)$ is the density of Q_ψ with respect to μ. The usual examples are $X = \mathbb{R}^d$, with $\mu = \lambda^d$, so $d\mu(x) = dx_1 dx_2 \cdots dx_d$, or X is a countable set with counting measure, with f a probability mass function or frequency function.

Bickel and Doksum give a slightly different but equivalent definition of exponential family, with $C(\theta) = e^{-B(\theta)}$, so that $B(\theta)$ is subtracted from the sum that is exponentiated.

Letting $T := T(x) := \{T_j(x)\}_{j=1}^k$ and $\theta := \theta(\psi)$, f can be written more briefly as $f = C(\theta)h(x)e^{\theta\cdot T}$ where $\theta \cdot T := \sum_{j=1}^k \theta_j T_j$. If we replace μ by ν where $d\nu(x) = h(x)d\mu(x)$, the factor $h(x)$ can be omitted, and ν is still a σ-finite measure. Given the θ_j, T_j, h, and μ, the number $C(\theta(\psi))$ is determined by normalization, so it is, in fact, a function of $\theta(\psi)$. Thus, given T_j, h, and μ, Q_ψ is determined by $\theta(\psi)$.

Example. The family of all normal laws $N(\mu, \sigma^2)$ on $\mathbb{R}$ for $-\infty < \mu < +\infty$ and $0 < \sigma < +\infty$ is an exponential family with $k = 2$, $\psi = (\mu, \sigma)$, $T_1(x) = x$, $\theta_1(\psi) = \mu/\sigma^2$, $T_2(x) = x^2$, $\theta_2(\psi) = -1/(2\sigma^2)$, and

$$C(\theta(\psi)) = (\sigma\sqrt{2\pi})^{-1}\exp(-\mu^2/(2\sigma)^2,$$

with $h(x) \equiv 1$.

It follows from the factorization theorem for sufficient statistics (in the "Sufficiency" handout) that for any exponential family, the vector-valued statistic $(T_1(x), \ldots, T_k(x))$ is a sufficient statistic. The structure of an exponential family is essentially preserved by taking n i.i.d. observations, since the following is clear:

Theorem 1. *Let $\{Q_\psi, \ \psi \in \Psi\}$ be any exponential family and let $X_1, \ldots, X_n$ be i.i.d. (Q_ψ). Then the distribution Q_ψ^n of $(X_1, \ldots, X_n)$ is an exponential family for the σ-finite measure μ^n on X^n, replacing $T_j(x)$ by $\sum_{i=1}^n T_j(X_i)$, $h(x)$ by $\Pi_{j=1}^n h(X_j)$, and $C(\theta(\psi))$ by $C(\theta(\psi))^n$. Thus for $(X_1, \ldots, X_n)$, the k-vector $\{\sum_{i=1}^n T_j(X_i)\}_{j=1}^k$ is a sufficient statistic.*

A family of $\mathcal{P}$ of probability distributions is called *equivalent* if for any P and Q in $\mathcal{P}$ with likelihood ratio $R_{Q/P}$, $0 < R_{Q/P} < +\infty$ (with probability 1 for P or Q). Since exponentials are strictly positive, any exponential family is equivalent.

The T_j will be called *affinely dependent* if for some constants c_0, c_1, $\ldots, c_k$, not all 0, $c_0 + c_1T_1 + \cdots + c_kT_k = 0$ almost everywhere for μ. Then $c_i \neq 0$ for some $i \geq 1$, and we can solve for T_i as a linear combination of other T_j and a constant. Then we can eliminate the T_i term and reduce k by 1, adding $-c_j\theta_i(\cdot)/c_i$ to each $\theta_j(\cdot)$ for $j \neq i$ and multiplying $C(\theta(\psi))$ by $\exp(-c_0\theta_i(\psi)/c_i)$. Iterating this, we can assume that $T_1, \ldots, T_k$ are affinely independent, i.e. they are not affinely dependent. Likewise, we can define affine independence for the functions θ_j, where now the linear relations among the $\theta_j(\cdot)$ and a constant would hold everywhere rather than almost everywhere (at this point we are not assuming a prior given on the parameter space Ψ). We can eliminate terms until $\theta_j(\cdot)$ are also affinely independent. We will always still have $k \geq 1$ since $\mathcal{P}$ contains at least two laws.

Let Θ be the range of the function $\psi \mapsto \theta(\psi) := (\theta_1(\psi), \ldots, \theta_k(\psi))$ from Ψ into $\mathbb{R}^k$:

$$\Theta := \left\{\theta(\psi) = \ \{\theta_j(\psi)\}_{j=1}^k : \ \psi \in \Psi\right\}. \tag{2}$$

Then clearly $\theta_1(\cdot), \ldots, \theta_k(\cdot)$ are affinely independent if and only if Θ is not included in any $(k-1)$-dimensional hyperplane in $\mathbb{R}^k$. Likewise, $T_1, \ldots, T_k$ are affinely independent (as defined above) if and only if for $T := (T_1, \ldots, T_k)$ from X into $\mathbb{R}^k$, the measure $\mu \circ T^{-1}$ is not concentrated in any $(k-1)$-dimensional hyperplane in $\mathbb{R}^k$.

A representation (1) of an exponential family will be called *minimal* if $T_1, \ldots, T_k$ are affinely independent, as are $\theta_1(\cdot), \ldots, \theta_k(\cdot)$.

A *functionoid* is an equivalence class of functions for the relation of almost everywhere equality for a measure. For functions on a Euclidean space $\mathbb{R}^n$, or an open subset of it, with Lebesgue measure λ^n, we will most often have statistics that are continuous functions of $x = (x_1, ..., x_n)$. In an equivalence class of functions for almost everywhere equality, there is at most one continuous function, and if we have continuous statistics, we can talk about functions rather than functionoids.

Any exponential family $\mathcal{P}$ as in (1) can be parameterized by the subset Θ of $\mathbb{R}^k$ given in (2), so that we get

$$(3) \qquad (dP_\theta/d\mu)(x) \;=\; C(\theta)h(x)\exp\left(\sum_{j=1}^{k}\theta_j T_j(x)\right), \quad \theta \in \Theta \subset \mathbb{R}^k,$$

where now $Q_\psi = P_{\theta(\psi)}$ for all $\psi \in \Psi$.

Theorem 2. *Every exponential family $\mathcal{P} := \{Q_\psi : \psi \in \Psi\}$ has a minimal representation (1), and then k is uniquely determined.*

Proof. We already saw that the $T_j(\cdot)$ can be taken to be affinely independent, as can the $\theta_j(\cdot)$, so that the representation (1) is minimal. Then in the representation (3), Θ, as mentioned, is not included in any $(k-1)$-dimensional hyperplane. The likelihood ratios are all of the form

$$R_{\theta,\phi} \;:=\; R_{P_\theta/P_\phi} \;=\; C(\theta)C(\phi)^{-1}\exp\left\{\sum_{j=1}^{k}(\theta_j - \phi_j)T_j(x)\right\}.$$

The logarithms of these likelihood ratios (log likelihood ratios) plus constants span a real vector space V_T of function(oid)s on X, included in the vector space W_T of function(oid)s spanned by $1, T_1, \ldots, T_k$. Then W_T is $(k+1)$-dimensional since $T_1, \ldots, T_k$ are affinely independent by minimality. Also, since $\theta_1, \ldots, \theta_k$ are affinely independent on Θ, $V_T = W_T$. Now $V := V_T$ is determined by the family $\mathcal{P}$, not depending on the choice of μ or T, so V and k are uniquely defined for the family $\mathcal{P}$. □

The number k is called here the *order* of the exponential family. Bickel and Doksum call it the *rank*. From here on it will be assumed that the representation of an exponential family is minimal unless it is specifically said not to be. The parameterization in (3) is then one-to-one:

Theorem 3. *If an exponential family has a minimal representation (3), then for any $\theta \neq \phi$ in Θ, $P_\theta \neq P_\phi$.*

Proof. If $P_\theta = P_\phi$, then for $\theta{\cdot}T \ :=\ \sum_j \theta_j T_j$, we have almost everywhere

$$\theta \cdot T + \log C(\theta) \ = \ \phi \cdot T + \log C(\phi),$$

or $(\theta - \phi) \cdot T = c$ for some c not depending on x. But $\theta \neq \phi$ means that the T_j are affinely dependent, contradicting minimality. □

Any subset of an exponential family is also an exponential family with the same T_j and ν, recalling that $d\nu(x) \ :=\ h(x)d\mu(x)$. It can be useful to take an exponential family as large as possible. Given ν and $T_j,\ j = 1, \ldots, k$ the *natural parameter space* of the exponential family is the set of all $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ such that

$$(4) \qquad K(\theta) \ :=\ \int \exp\left(\sum_{j=1}^{k} \theta_j T_j(x)\right) d\nu(x) \ <\ \infty.$$

Clearly $K(\theta) > 0$ for all θ. For any θ in the natural parameter space, we can define $C(\theta) \ :=\ 1/K(\theta)$ and get a probability measure P_θ given by (3). So we have a family of laws P_θ indexed by the natural parameter space. The family doesn't extend to values of θ outside the natural parameter space since then normalization is not possible.

Theorem 4. *For any given σ-finite ν and measurable functions T_j on $(X, \mathcal{B})$, the natural parameter space is a convex set in $\mathbb{R}^k$.*

Proof. First, for any real y (which can be positive or negative), $\theta \mapsto e^{y\theta}$ is a convex function of $\theta \in \mathbb{R}$ (its second derivative is positive, so its first derivative is increasing, which implies convexity). It follows that for any real $y_1, \ldots, y_k$, the function

$$\theta \ = \ (\theta_1, \ldots, \theta_k) \mapsto \exp(y_1\theta_1 + \cdots + y_k\theta_k)$$

is convex on $\mathbb{R}^k$. The inequalities defining convexity are preserved when integrated with respect to a nonnegative measure, so $K(\theta)$ is a convex function, whose values may be infinite for some θ (just those θ outside the natural parameter space). The set where a convex function $< +\infty$ is clearly a convex set. □

The usual theorem on interchanging integrals for integrable functions, or nonnegative measurable functions, of two or more variables, is usually called the Fubini theorem. I call it the Tonelli–Fubini theorem because Fubini first stated it, but Tonelli first proved it correctly.

Proposition 5. *For any exponential family, the natural parameter space is the same for any number n of i.i.d. observations.*

Proof. If $K_n(\theta)$ is the integral $K(\theta)$ for n observations, then from the definitions and the Tonelli–Fubini theorem, $K_n(\theta) = K_1(\theta)^n$ for all n, so $K_n(\theta)$ is finite if and only if $K_1(\theta)$ is. □

Theorem 6. *For an exponential family as in (3) let U be the interior of the natural parameter space. Then for $\xi = (\xi_1, \ldots, \xi_k)$ in U and $\eta = (\eta_1, \ldots, \eta_k) \in \mathbb{R}^k$, let $W := \{\zeta = \xi + i\eta : \ \xi \in U, \ \eta \in \mathbb{R}^k\}$ so that $\zeta_j = \xi_j + i\eta_j$ for $j = 1, \ldots, k$. Then the function $K(z)$ in (4) is, on W, an analytic (holomorphic) function of z, representable by a power series in the k coordinates $z_j - \zeta_j$ in the neighborhood of any point ζ in W. In particular K has, on W, continuous partial derivatives of all orders with respect to z, which can be obtained by differentiating under the integral sign. In other words, for any $p = (p_1, \ldots, p_k)$, where the $p(i) := p_i$ are nonnegative integers and $[p] := p_1 + \cdots p_k$, the partial derivative $D^pK := \partial^{[p]}K(z)/\partial z_1^{p(1)} \cdots \partial z_k^{p(k)}$ exists and is continuous, and equals $\int T(x)^p \exp(\sum_{j=1}^k z_j T_j(x)) d\nu(x)$, where $t^p := t_1^{p(1)} \cdots t_k^{p(k)}$. For any $\xi \in U$, $E_\xi T^p = D^pK(\xi)/K(\xi)$.*

Proof. Let $\zeta = \xi + i\eta \in W$, so $\xi \in U$ and $\eta \in \mathbb{R}^k$. Take $\varepsilon > 0$ small enough so that if $|u_j - \xi_j| \leq \varepsilon$ for all $j = 1, \ldots, k$ then $u \in U$, so $u + iv \in W$ for any $v \in \mathbb{R}^k$. Then for any $T = T(x) \in \mathbb{R}^k$,

$$|e^{(u+iv)\cdot T}| \ = \ e^{u\cdot T} \ = \ e^{(u-\xi)\cdot T} e^{\xi\cdot T}.$$

Thus, replacing $d\nu(x)$ by $e^{\xi\cdot T(x)}d\nu(x)$, we can assume that $\xi = 0$. Then $|u_j| \leq \varepsilon$ for $j = 1, \ldots, k$.

We have $e^{u\cdot T} \ = \ \Pi_{j=1}^k \exp(u_j T_j)$,

$$\exp(u_1T_1) \quad = \quad \sum_{r=0}^{\infty} (u_1T_1)^r/r!, \quad |(u_1T_1)^r| \ = \ |u_1|^r |T_1|^r,$$

and

$$\sum_{r=0}^{\infty} |u_1T_1|^r/r! \quad = \quad \exp(|u_1T_1|) \ \leq \ \exp(-\varepsilon T_1) + \exp(\varepsilon T_1),$$

and likewise for any $j = 2, \ldots, k$ in place of $j = 1$. By choice of ε,

$$\int \cdots \int \Pi_{j=1}^k \exp(\pm\varepsilon T_j) d\nu(x_1) \cdots d\nu(x_k) < \infty$$

for any choices of ±, where $T_j := T_j(x_j)$ for each j, so the sum over all 2^k possible choices of ± of the integrals is finite. Thus by dominated convergence, the series

$$e^{u\cdot T} \ = \ \Pi_{j=1}^k \sum_{r_j=0}^{\infty} (u_jT_j)^{r_j}/r_j!$$

converges absolutely if $|u_j| \le \varepsilon$ for all j, and can be interchanged with

$$\int\int\cdots\int \cdot d\nu(x_1)\cdots d\nu(x_k).$$

The integral yields a power series in $u_1, \ldots, u_k$. In the above, u_j can be replaced by $u_j + iv_j$ if $|u_j + iv_j| \le \varepsilon$ for each j. So we get a power series converging to $K(z)$ for $z = u + iv$. Since such a series exists in some neighborhood of each point in W, $K(\cdot)$ is holomorphic on W as stated.

To show that derivatives can be taken under the integral sign, first let $k = 1$ and $p = 1$. If $0 < t < c$ and $y > 0$ then for $\lambda := t/c$ and $x := cy$, by convexity $e^{\lambda x} \le \lambda e^x + (1-\lambda)e^0 \le \lambda e^x + 1$, so $(e^{ty} - 1)/t \le e^{cy}/c$. Likewise, for $0 < |t| < c$ and all y, $|(e^{ty} - 1)/t| \le (e^{cy} + e^{-cy})/c$. For u in U, and c small enough, $u \pm c \in U$, so the functions $\{(e^{(t+u)T(x)} - e^{uT(x)})/t : \ 0 < |t| < c\}$ are dominated by an integrable function. So

$$\frac{d}{d\theta}\int e^{\theta T(x)} d\nu(x)|_{\theta=u} \ = \ \int T(x)e^{uT(x)} d\nu(x).$$

Also, $|y| \le (e^{cy} + e^{-cy})/c$.

For $p > 1$, $c^{-p}(e^{cy} + e^{-cy})^p \le (2/c)^p(e^{pcy} + e^{-pcy})$. For fixed p, $u \pm pc \in U$ for c small enough, so we can again apply dominated convergence to get

$$(d^p/d\theta^p)\int e^{\theta T(x)} d\nu(x)|_{\theta=u} \ = \ \int T(x)^p e^{uT(x)} d\nu(x).$$

Now for $k > 1$, and any $p \in \mathbb{N}^k$, the 2^k (or fewer) points
$(u_1 \pm cp_1, \ldots, u_k \pm cp_k)$ are all in U if c is small enough. Dominated convergence applies once more, so the derivatives can be interchanged with integrals as stated.

The final statement follows easily since $C(\theta) \equiv 1/K(\theta)$, finishing the proof. □

Suppose given an exponential family as in (3) and let $j(\theta) := \log K(\theta) = -\log C(\theta)$, so that $dP_\theta/d\nu = \exp(-j(\theta) + \theta\cdot T)$ where $\theta \cdot T := \sum_j \theta_j T_j(x)$. Since the vector $T := \{T_j\}_{j=1}^k$ gives a sufficient statistic for the family, the means and variances of its components are of interest. They have nice expressions in terms of derivatives of the function j. The gradient of j is the vector-valued function $\triangledown j := (\partial j/\partial\theta_1, \ldots, \partial j/\partial\theta_k)$.

Corollary 7. *Suppose given an exponential family of order k in minimal form (3). Then the natural parameter space Θ has non-empty interior $U \subset \mathbb{R}^k$. For any $\theta \in U$, $E_\theta T = \triangledown j(\theta)$ and for any $r, s = 1, ..., k$,*

$$(5) \qquad \mathrm{Cov}_\theta(T_r, T_s) = E_\theta(T_r T_s) - E_\theta T_r E_\theta T_s = \partial^2 j(\theta)/\partial\theta_r \partial\theta_s.$$

On U, j is a strictly convex function.

Proof. Any convex set in $\mathbb{R}^k$ either has non-empty interior or is included in some lower-dimensional affine subspace (Dudley [2002, 6.2.6]). The latter would imply that $\theta_1, ..., \theta_k$ are affinely dependent, contradicting the minimal form. So Θ has a non-empty interior U as stated.

Theorem 6 gives $E_\theta(T_r T_s) = (\partial^2 K/\partial\theta_r\partial\theta_s)/K(\theta)$ and

$$E_\theta T_r = (\partial K/\partial\theta_r)/K(\theta) = \partial j(\theta)/\partial\theta_r.$$

This implies that $E_\theta T = \triangledown j(\theta)$. Taking $\partial/\partial\theta_s$ of both sides of the latter equation in the last display gives (5).

Any covariance matrix is symmetric and nonnegative definite. To show $\mathrm{Cov}_\theta(T_r, T_s)$ is positive definite, suppose not. Then for some $a_1, ..., a_k$ not all 0,

$$0 = \sum_{r,s=1}^{k} \mathrm{Cov}_\theta(T_r, T_s) a_r a_s = \mathrm{Var}_\theta \left(\sum_{r=1}^{k} a_r T_r \right).$$

Then $\sum_{r=1}^k a_r T_r$ equals a constant almost surely, contradicting affine independence of $T_1, ..., T_k$ (minimal form). So the Hessian matrix of second partial derivatives of j is positive definite. Now consider j along a line segment included in U, $(1-\lambda)s + \lambda t = s + \lambda(t - s)$ for $s \neq t$ in U and $0 \leq \lambda \leq 1$. From the chain rule we get that $\partial^2 j(s+\lambda(t-s))/\partial\lambda^2 > 0$ for $0 < \lambda < 1$. It's easily seen that a function of a real variable on an open interval with a strictly positive second derivative is strictly convex. This implies (since j is smooth on U by Theorem 6) that j is strictly convex on U, proving the Corollary. □

Next is a description of posterior distributions for exponential families. It follows from a fact about sufficiency of sufficient statistics for posterior distributions, given in "Sufficiency."

Theorem 8. *Suppose given an exponential family with likelihood function $f(\theta, x) = C(\theta)h(x)e^{\theta \cdot T(x)}$ as in (3), in minimal form, of order k, for $\theta \in \Theta$ where Θ is the natural parameter space. Let π be any prior on Θ. Then for any x with $h(x) > 0$,*

(a) *π_x has a density with respect to π given by*

$$(6) \qquad \frac{d\pi_x}{d\pi}(\theta) = \frac{C(\theta)\exp(\theta \cdot T(x))}{\int C(\phi)\exp(\phi \cdot T(x)) d\pi(\phi)}.$$

(b) *If π has a density Π with respect to Lebesgue measure λ^k on Θ, then we also have*

$$\frac{d\pi_x}{d\lambda^k}(\theta) = \frac{\Pi(\theta)C(\theta)\exp(\theta \cdot T(x))}{\int \Pi(\phi)C(\phi)\exp(\phi \cdot T(x))d\lambda^k(\phi)}. \tag{7}$$

(c) *If we have n i.i.d. observations $X_1, ..., X_n$ and $x = (X_1, ..., X_n)$, the above equations hold with $C(\psi)$ replaced by $C(\psi)^n$ for $\psi = \theta$ or ϕ, and $T(x)$ by $\sum_{j=1}^n T(X_j) \in \mathbb{R}^k$.*

The existence of a k-dimensional sufficient statistic $T = (T_1, \ldots, T_k)$ for an exponential family extends to any sample size n for n i.i.d. observations, as noted previously, replacing each T_i by $\sum_{j=1}^n T_i(X_j)$. When R. A. Fisher first defined exponential families, one of the main properties he pointed out was the possibility of data reduction in this way. Moreover, he stated that if the data can be reduced, in other words if for i.i.d. $X_1, \ldots, X_n$ there is a sufficient statistic of dimension $k < n$ (even for one value of n) then the family of laws must be exponential. This is true under some regularity conditions, one of which is that the family be equivalent. For example, the family of uniform distributions on intervals $[0, \theta]$, $0 < \theta < \infty$, has a 1-dimensional sufficient statistic, the largest order statistic $X_{(n)}$, but is evidently not equivalent and (so) not exponential. Other regularity conditions of continuity and differentiability will be assumed. If there were no such conditions, the "dimension" of a sufficient statistic would not be meaningful. For example, if X and Y are any two uncountable Borel sets in complete separable metric spaces, such as $X = \mathbb{R}^k$ and $Y = \mathbb{R}^m$, then there is always a 1-1, Borel measurable function from X onto Y with measurable inverse (Dudley [2002, Section 13.1]). Any Borel measurable function is continuous when restricted to sets having nearly full measure (Lusin's theorem, Dudley [2002, Theorem 7.5.2]). Also, for any m there is a continuous function from $\mathbb{R}^m$ into $\mathbb{R}$, 1-1 almost everywhere for Lebesgue measure (Denny, 1964).

The following example may illustrate the point. Let x and y be two numbers in $[0, 1]$, each represented by its decimal expansion, $x = \sum_{n\geq 1} x_n/10^n$ where each x_n is $0, 1, \ldots,$ or 9, and likewise for y. By alternating digits define a real number z with digits $z_{2n-1} = x_n$ and $z_{2n} = y_n$ for $n = 1, 2, \ldots$. This gives a correspondence between ordered pairs (x, y) of real numbers and individual real numbers z. Although it is not quite well-defined, because of ambiguities such as $0.099999999\ldots = 0.100000\ldots$, 1-1 or continuous, the correspondence illustrates a reduction of dimension (from 2 to 1) which is not a real reduction in the sense of statistical interest. The example also shows why some regularity conditions such as differentiability may be expected in proofs about data reduction implying that a family is exponential.

Let $\mathcal{P}$ be an equivalent family of probability measures. Let Q be a fixed law in the family. If T is a sufficient statistic for $\{P^n : P \in \mathcal{P}\}$, the family of laws of n i.i.d. observations $X_1, \ldots, X_n$ with laws in $\mathcal{P}$, then by the factorization theorem, for each P in $\mathcal{P}$ there is a function ρ_P with

$$(8) \qquad \Pi_{j=1}^n R_{P/Q}(x_j) \;=\; \rho_P(T(x_1, \ldots, x_n))$$

for almost all $x_1, \ldots, x_n$. T will be called *strongly sufficient* (with respect to given choices of Q and of $R_{P/Q}$ for all x and all $P \in \mathcal{P}$) if (8) holds for *all* (and not only almost all) x.

Let $\phi_P(x) \;:=\; \log R_{P/Q}(x)$ for any P in $\mathcal{P}$. We will be considering families for which the likelihood ratios $R_{P/Q}$ are continuous non-zero functions of x, so that ϕ_P is continuous, and where every neighborhood of each point in the sample space has positive measure for each law in $\mathcal{P}$, so that ϕ_P is determined everywhere by continuity and not only almost everywhere. So, strong sufficiency is a reasonable assumption.

A function f on a region in $\mathbb{R}^k$ is called C^1 if it has continuous first partial derivatives with respect to each of the k variables. It will be called BC^1 if these derivatives are also bounded. A real-valued function f on an interval $U \subset \mathbb{R}$ will be called *piecewise* BC^1 if f is continuous on U and there is a finite set $F \subset U$ such that f is BC^1 on $U \setminus F$, i.e. f is BC^1 on each of finitely many open intervals whose endpoints are in F or are endpoints of U. Now a fact can be stated:

Theorem 9. *Let $\mathcal{P}$ be a family of laws defined on a connected open set U in $\mathbb{R}^r$ and having densities f_P, $P \in \mathcal{P}$, with respect to Lebesgue measure λ on U, with $f_P(x) > 0$ for all $x \in U$ and P in $\mathcal{P}$ (so $\mathcal{P}$ is equivalent). Suppose that all the functions f_P are continuous on U and that for some positive integers $k < n$, there is a statistic T, continuous from U^n into $\mathbb{R}^k$, strongly sufficient for $\{P^n : P \in \mathcal{P}\}$, where $R_{P/Q} \;:=\; f_P/f_Q$. Then*

(a) *If $k = 1$, $\mathcal{P}$ is an exponential family of order 1.*
(b) *If all the densities f_P are BC^1, or if $r = 1$ and they are piecewise BC^1, then $\mathcal{P}$ is exponential of order at most k.*

The proof is too long to be given here. It appears in the 18.466 OCW notes, Theorem 2.5.11.

Example. This will show why the connectedness of U is, or the continuity hypotheses are, needed in Theorem 9. Let $U \;:=\; (0,1) \cup (2,3)$ (which is not connected). Let the dominating measure ν be the sum of

Lebesgue measures on the two intervals. For $0 < \lambda < 1, 0 < \theta < \infty$ let

$$\begin{aligned} \phi_{\theta,\lambda}(x) &:= \lambda\theta e^{\theta x}/(e^\theta - 1), && 0 < x < 1; \\ &:= (1-\lambda)\theta e^{\theta(x-2)}/(e^\theta - 1), && 2 < x < 3. \end{aligned}$$

It is straightforward to check that this is a probability density for each θ and λ. Let x_1, x_2 be i.i.d. with this density. Then the likelihood function is

$$u(\theta, \lambda, x_1, x_2)\theta^2 \exp(\theta(x_1 + x_2))/(e^\theta - 1)^2$$

where $u(\theta, \lambda, x_1, x_2) := \lambda^2$ for $0 < x_1 + x_2 < 2$, $2\lambda(1-\lambda)e^{-2\theta}$ for $2 < x_1 + x_2 < 4$, and $(1-\lambda)^2 e^{-4\theta}$ for $4 < x_1 + x_2 < 6$. It follows by the factorization theorem, not only because of the factor $\exp(\theta(x_1 + x_2))$ but because the ranges for different formulas for $u(\cdot, \cdot, x_1, x_2)$ also are functions of $x_1 + x_2$, that $x_1 + x_2$ is a $k = 1$-dimensional sufficient statistic for the family with $n = 2$. Let $\gamma(\theta, \lambda) := \log[\theta/(e^\theta - 1)]$. Then one can check that

$$\log \phi_{\theta,\lambda}(x) = \gamma(\theta, \lambda) + \log \lambda + \theta x + [\log((1-\lambda)/\lambda) - 2\theta]1_{2<x<3}.$$

Since the functions x and $1_{2<x<3}$ are affinely independent, as are the functions θ and $\log((1-\lambda)/\lambda)$, we see that the family is exponential of order 2, not 1. Thus the conclusion of Theorem 9(a) does not hold in this case. The connectedness of the interval U is used in the proof more than once, by way of the intermediate value theorem.

Of course, connectedness is only meaningful in connection with continuity of some functions. We could take $U := (0, 2)$ to be connected in the example while $\phi_{\theta,\lambda}$ and T are discontinuous by replacing $(2, 3)$ by $[1, 2)$ and letting $x_1(x) = x$ for $0 < x < 1$ and $x_1(x) = x + 1$ for $1 \le x < 2$, while taking x_2 as an i.i.d. copy of x_1.

Or, replacing the union of two intervals by a union of as many intervals as we like, we can get the exponential family to be of arbitrarily high order for $n = 2$. Similarly, by spreading the intervals farther apart, for example taking $(0, 1) \cup (n, n + 1) \cup (n^2 + n, n^2 + n + 1), \ldots$, we can get a 1-dimensional sufficient statistic for any number n of i.i.d. observations, again if U is not connected or the densities and T are not continuous.

Note: for $r > 1$, the dimension of the full data vector $(X_1, \ldots, X_n)$ is nr, but that the assumption in Theorem 9 is $k < n$ (and not $k < nr$). Suppose we consider a family of distributions $\mathbb{R}^r$ having densities with respect to some measure (not Lebesgue measure) which are functions of the first coordinate x_1. Then x_1 is a sufficient statistic. For n i.i.d. variables there is an n-dimensional sufficient statistic and $n < nr$, but the family need not be exponential. So the assumption $k < n$ in Theorem 9 is sharp.

Notes. Two classic books on exponential families are Barndorff-Nielsen (1978) and Brown (1986). There are more recent books, which I have not seen, on particular aspects, such as *Exponential Families of Stochastic Processes* (Uwe Küchler et al., 1997) and *Exponential Family Nonlinear Models* (Bo-Cheng Wei, 1998).

REFERENCES

Barndorff-Nielsen, O. (1978). *Information and Exponential Families.* Wiley, New York.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, IMS Lecture Notes–Monograph Series **9**, 283 pp.