MATHEMATICAL STATISTICS

BASIC IDEAS AND SELECTED TOPICS

VOL. I

SECOND EDITION

PETER J. BICKEL

KJELL A. DOKSUM

observations $\mathbf{X}$ to be the function of θ defined by $L_{\mathbf{X}}(\theta) = p(\mathbf{X}, \theta)$, $\theta \in \Theta$. If $T(\mathbf{X})$ is sufficient for θ, and if there is a value $\theta_0 \in \Theta$ such that

$$\{\mathbf{x} : p(\mathbf{x}, \theta) > 0\} \subset \{\mathbf{x} : p(\mathbf{x}, \theta_0) > 0\},\ \theta \in \Theta,$$

then, by the factorization theorem, the *likelihood ratio*

$$\Lambda_{\mathbf{X}}(\theta) = \frac{L_{\mathbf{X}}(\theta)}{L_{\mathbf{X}}(\theta_0)}$$

depends on $\mathbf{X}$ through $T(\mathbf{X})$ only. $\Lambda_{\mathbf{X}}(\theta)$ is a minimally sufficient statistic.

1.6 EXPONENTIAL FAMILIES

The binomial and normal models considered in the last section exhibit the interesting feature that there is a natural sufficient statistic whose dimension as a random vector is independent of the sample size. The class of families of distributions that we introduce in this section was first discovered in statistics independently by Koopman, Pitman, and Darmois through investigations of this property[1]. Subsequently, many other common features of these families were discovered and they have become important in much of the modern theory of statistics.

Probability models with these common features include normal, binomial, Poisson, gamma, beta, and multinomial regression models used to relate a response variable Y to a set of predictor variables. More generally, these families form the basis for an important class of models called generalized linear models. We return to these models in Chapter 2. They will reappear in several connections in this book.

1.6.1 The One-Parameter Case

The family of distributions of a model $\{P_\theta : \theta \in \Theta\}$, is said to be a *one-parameter exponential family*, if there exist real-valued functions $\eta(\theta)$, $B(\theta)$ on Θ, real-valued functions T and h on R^q, such that the density (frequency) functions $p(x, \theta)$ of the P_θ may be written

$$p(x, \theta) = h(x) \exp\{\eta(\theta)T(x) - B(\theta)\} \tag{1.6.1}$$

where $x \in \mathcal{X} \subset R^q$. Note that the functions η, B, and T are not unique.

In a one-parameter exponential family the random variable $T(X)$ is sufficient for θ. This is clear because we need only identify $\exp\{\eta(\theta)T(x) - B(\theta)\}$ with $g(T(x), \theta)$ and $h(x)$ with itself in the factorization theorem. We shall refer to T as a *natural sufficient statistic* of the family.

Here are some examples.

Example 1.6.1. *The Poisson Distribution.* Let P_θ be the Poisson distribution with unknown mean θ. Then, for $x \in \{0, 1, 2, \cdots\}$,

$$p(x, \theta) = \frac{\theta^x e^{-\theta}}{x!} = \frac{1}{x!} \exp\{x \log \theta - \theta\},\ \theta > 0. \tag{1.6.2}$$

Therefore, the P_θ form a one-parameter exponential family with

$$q = 1, \eta(\theta) = \log\theta, B(\theta) = \theta, T(x) = x, h(x) = \frac{1}{x!}. \tag{1.6.3}$$

Example 1.6.2. *The Binomial Family.* Suppose X has a $\mathcal{B}(n,\theta)$ distribution, $0 < \theta < 1$. Then, for $x \in \{0, 1, \ldots, n\}$

$$\begin{aligned} p(x,\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}\exp[x\log(\frac{\theta}{1-\theta}) + n\log(1-\theta)]. \end{aligned} \tag{1.6.4}$$

Therefore, the family of distributions of X is a one-parameter exponential family with

$$q = 1, \eta(\theta) = \log(\frac{\theta}{1-\theta}), B(\theta) = -n\log(1-\theta), T(x) = x, h(x) = \binom{n}{x}. \tag{1.6.5}$$

□

Here is an example where $q = 2$.

Example 1.6.3. Suppose $X = (Z, Y)^T$ where $Y = Z + \theta W$, $\theta > 0$, Z and W are independent $\mathcal{N}(0,1)$. Then

$$\begin{aligned} f(x,\theta) &= f(z,y,\theta) = f(z)f_\theta(y \mid z) = \varphi(z)\theta^{-1}\varphi((y-z)\theta^{-1}) \\ &= (2\pi\theta)^{-1}\exp\left\{-\frac{1}{2}[z^2 + (y-z)^2\theta^{-2}]\right\} \\ &= (2\pi)^{-1}\exp\left\{-\frac{1}{2}z^2\right\}\exp\left\{-\frac{1}{2}\theta^{-2}(y-z)^2 - \log\theta\right\}. \end{aligned}$$

This is a one-parameter exponential family distribution with

$$q = 2, \eta(\theta) = -\frac{1}{2}\theta^{-2}, B(\theta) = \log\theta, T(x) = (y-z)^2, h(x) = (2\pi)^{-1}\exp\left\{-\frac{1}{2}z^2\right\}.$$

□

The families of distributions obtained by sampling from one-parameter exponential families are themselves one-parameter exponential families. Specifically, suppose $X_1, \ldots, X_m$ are independent and identically distributed with common distribution P_θ, where the P_θ form a one-parameter exponential family as in (1.6.1). If $\{P_\theta^{(m)}\}$, $\theta \in \Theta$, is the family of distributions of $\mathbf{X} = (X_1, \ldots, X_m)$ considered as a random vector in R^{mq} and $p(\mathbf{x},\theta)$ are the corresponding density (frequency) functions, we have

$$\begin{aligned} p(\mathbf{x},\theta) &= \prod_{i=1}^m h(x_i)\exp[\eta(\theta)T(x_i) - B(\theta)] \\ &= \left[\prod_{i=1}^m h(x_i)\right]\exp\left[\eta(\theta)\sum_{i=1}^m T(x_i) - mB(\theta)\right] \end{aligned} \tag{1.6.6}$$

where $\mathbf{x} = (x_1, \ldots, x_m)$. Therefore, the $P_\theta^{(m)}$ form a one-parameter exponential family. If we use the superscript m to denote the corresponding T, η, B, and h, then $q^{(m)} = mq$, and

$$\begin{aligned} \eta^{(m)}(\theta) &= \eta(\theta), \\ T^{(m)}(\mathbf{x}) &= \sum_{i=1}^{m} T(x_i), B^{(m)}(\theta) = mB(\theta), h^{(m)}(\mathbf{x}) = \prod_{i=1}^{m} h(x_i). \end{aligned} \tag{1.6.7}$$

Note that the natural sufficient statistic $T^{(m)}$ is one-dimensional whatever be m. For example, if $\mathbf{X} = (X_1, \ldots, X_m)$ is a vector of independent and identically distributed $\mathcal{P}(\theta)$ random variables and $P_\theta^{(m)}$ is the family of distributions of $\mathbf{x}$, then the $P_\theta^{(m)}$ form a one-parameter exponential family with natural sufficient statistic $T^{(m)}(\mathbf{X}) = \sum_{i=1}^{m} X_i$.

Some other important examples are summarized in the following table. We leave the proof of these assertions to the reader.

TABLE 1.6.1

Family of distributions		$\eta(\theta)$	$T(x)$
$\mathcal{N}(\mu, \sigma^2)$	σ^2 fixed	μ/σ^2	x
	μ fixed	$-1/2\sigma^2$	$(x-\mu)^2$
$\Gamma(p, \lambda)$	p fixed	$-\lambda$	x
	λ fixed	$(p-1)$	$\log x$
$\beta(r, s)$	r fixed	$(s-1)$	$\log(1-x)$
	s fixed	$(r-1)$	$\log x$

The statistic $T^{(m)}(X_1, \ldots, X_m)$ corresponding to the one-parameter exponential family of distributions of a sample from any of the foregoing is just $\sum_{i=1}^{m} T(X_i)$.

In our first Example 1.6.1 the sufficient statistic $T^{(m)}(X_1, \ldots, X_m) = \sum_{i=1}^{m} X_i$ is distributed as $\mathcal{P}(m\theta)$. This family of Poisson distributions is one-parameter exponential whatever be m. In the discrete case we can establish the following general result.

Theorem 1.6.1. Let $\{P_\theta\}$ be a one-parameter exponential family of discrete distributions with corresponding functions T, η, B, and h, then the family of distributions of the statistic $T(X)$ is a one-parameter exponential family of discrete distributions whose frequency functions may be written

$$h^*(t) \exp\{\eta(\theta)t - B(\theta)\}$$

for suitable h^*.

Proof. By definition,

$$\begin{aligned} P_\theta[T(x) = t] &= \sum_{\{x:T(x)=t\}} p(x,\theta) \\ &= \sum_{\{x:T(x)=t\}} h(x)\exp[\eta(\theta)T(x) - B(\theta)] \\ &= \exp[\eta(\theta)t - B(\theta)]\{\sum_{\{x:T(x)=t\}} h(x)\}. \end{aligned} \tag{1.6.8}$$

If we let $h^*(t) = \sum_{\{x:T(x)=t\}} h(x)$, the result follows. □

A similar theorem holds in the continuous case if the distributions of $T(X)$ are themselves continuous.

Canonical exponential families. We obtain an important and useful reparametrization of the exponential family (1.6.1) by letting the model be indexed by η rather than θ. The exponential family then has the form

$$q(x,\eta) = h(x)\exp[\eta T(x) - A(\eta)],\ x \in \mathcal{X} \subset R^q \tag{1.6.9}$$

where $A(\eta) = \log \int \cdots \int h(x)\exp[\eta T(x)]dx$ in the continuous case and the integral is replaced by a sum in the discrete case. If $\theta \in \Theta$, then $A(\eta)$ must be finite, if q is definable. Let $\mathcal{E}$ be the collection of all η such that $A(\eta)$ is finite. Then as we show in Section 1.6.2, $\mathcal{E}$ is either an interval or all of R and the class of models (1.6.9) with $\eta \in \mathcal{E}$ contains the class of models with $\theta \in \Theta$. The model given by (1.6.9) with η ranging over $\mathcal{E}$ is called the *canonical one-parameter exponential family generated by T and h*. $\mathcal{E}$ is called the *natural parameter space* and T is called the *natural sufficient statistic*.

Example 1.6.1. *(continued).* The Poisson family in canonical form is

$$q(x,\eta) = (1/x!)\exp\{\eta x - \exp[\eta]\},\ x \in \{0,1,2,\dots\},$$

where $\eta = \log\theta$,

$$\exp\{A(\eta)\} = \sum_{x=0}^{\infty}(e^{\eta x}/x!) = \sum_{x=0}^{\infty}(e^{\eta})^x/x! = \exp(e^{\eta}),$$

and $\mathcal{E} = R$.

Here is a useful result. □

Theorem 1.6.2. *If X is distributed according to (1.6.9) and η is an interior point of $\mathcal{E}$, the moment-generating function of $T(X)$ exists and is given by*

$$M(s) = \exp[A(s+\eta) - A(\eta)]$$

for s in some neighborhood of 0.

Moreover,

$$E(T(X)) = A'(\eta), \quad \mathrm{Var}(T(X)) = A''(\eta).$$

Proof. We give the proof in the continuous case. We compute

$$\begin{aligned} M(s) &= E(\exp(sT(X))) = \int \cdots \int h(x) exp[(s+\eta)T(x) - A(\eta)]dx \\ &= \{exp[A(s+\eta) - A(\eta)]\} \int \cdots \int h(x) exp[(s+\eta)T(x) - A(s+\eta)]dx \\ &= exp[A(s+\eta) - A(\eta)] \end{aligned}$$

because the last factor, being the integral of a density, is one. The rest of the theorem follows from the moment-enerating property of $M(s)$ (see Section A.12). □

Here is a typical application of this result.

Example 1.6.4 Suppose $X_1, \ldots, X_n$ is a sample from a population with density

$$p(x, \theta) = (x/\theta^2) \exp(-x^2/2\theta^2), \quad x > 0, \theta > 0.$$

This is known as the *Rayleigh* distribution. It is used to model the density of "time until failure" for certain types of equipment. Now

$$\begin{aligned} p(\mathbf{x}, \theta) &= (\prod_{i=1}^{n} (x_i/\theta^2)) \exp(-\sum_{i=1}^{n} x_i^2/2\theta^2) \\ &= (\prod_{i=1}^{n} x_i) \exp[\frac{-1}{2\theta^2} \sum_{i=1}^{n} x_i^2 - n \log \theta^2]. \end{aligned}$$

Here $\eta = -1/2\theta^2$, $\theta^2 = -1/2\eta$, $B(\theta) = n \log \theta^2$ and $A(\eta) = -n \log(-2\eta)$. Therefore, the natural sufficient statistic $\sum_{i=1}^{n} X_i^2$ has mean $-n/\eta = 2n\theta^2$ and variance $n/\eta^2 = 4n\theta^4$. Direct computation of these moments is more complicated. □

1.6.2 The Multiparameter Case

Our discussion of the "natural form" suggests that one-parameter exponential families are naturally indexed by a one-dimensional real parameter η and admit a one-dimensional sufficient statistic $T(x)$. More generally, Koopman, Pitman, and Darmois were led in their investigations to the following family of distributions, which is naturally indexed by a k-dimensional parameter and admit a k-dimensional sufficient statistic.

A family of distributions $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset R^k$, is said to be a *k-parameter exponential family*, if there exist real-valued functions $\eta_1, \ldots, \eta_k$ and B of $\boldsymbol{\theta}$, and real-valued functions $T_1, \ldots, T_k$, h on R^q such that the density (frequency) functions of the P_θ may be written as,

$$p(x, \boldsymbol{\theta}) = h(x) \exp[\sum_{j=1}^{k} \eta_j(\boldsymbol{\theta}) T_j(x) - \boldsymbol{B}(\boldsymbol{\theta})], \quad x \in \mathcal{X} \subset R^q. \tag{1.6.10}$$

By Theorem 1.3.1, the vector $\mathbf{T}(X) = (T_1(X), \ldots, T_k(X))^T$ is sufficient. It will be referred to as a *natural sufficient statistic* of the family.

Again, suppose $\mathbf{X} = (X_1, \ldots, X_m)$ where the X_i are independent and identically distributed and their common distribution ranges over a k-parameter exponential family given by (1.6.10). Then the distributions of $\mathbf{X}$ form a k-parameter exponential family with natural sufficient statistic

$$\mathbf{T}^{(m)}(x) = (\sum_{i=1}^{m} T_1(X_i), \ldots, \sum_{i=1}^{m} T_k(X_i)).$$

Example 1.6.5. *The Normal Family.* Suppose that $P_\theta = \mathcal{N}(\mu, \sigma^2)$, $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$. The density of P_θ may be written as

$$p(x, \boldsymbol{\theta}) = \exp[\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{1}{2}(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2))], \tag{1.6.11}$$

which corresponds to a two-parameter exponential family with $q = 1$, $\theta_1 = \mu$, $\theta_2 = \sigma^2$, and

$$\begin{aligned} \eta_1(\boldsymbol{\theta}) &= \frac{\mu}{\sigma^2},\ T_1(x) = x,\ \eta_2(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2},\ T_2(x) = x^2, \\ B(\boldsymbol{\theta}) &= \frac{1}{2}(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2)),\ h(x) = 1. \end{aligned}$$

If we observe a sample $\mathbf{X} = (X_1, \ldots, X_m)$ from a $\mathcal{N}(\mu, \sigma^2)$ population, then the preceding discussion leads us to the natural sufficient statistic

$$(\sum_{i=1}^{m} X_i, \sum_{i=1}^{m} X_i^2),$$

which we obtained in the previous section (Example 1.5.4). □

Again it will be convenient to consider the "biggest" families, letting the model be indexed by $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)^T$ rather than $\boldsymbol{\theta}$. Thus, *the canonical k-parameter exponential family generated by $\mathbf{T}$ and h* is

$$q(x, \boldsymbol{\eta}) = h(x) \exp\{\mathbf{T}^T(x)\boldsymbol{\eta} - A(\boldsymbol{\eta})\}, x \in \mathcal{X} \subset R^q$$

where $\mathbf{T}(x) = (T_1(x), \ldots, T_k(x))^T$ and, in the continuous case,

$$A(\boldsymbol{\eta}) = \log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x) \exp\{\mathbf{T}^T(x)\boldsymbol{\eta}\} dx.$$

In the discrete case, $A(\boldsymbol{\eta})$ is defined in the same way except integrals over R^q are replaced by sums. In either case, we define the natural parameter space as

$$\mathcal{E} = \{\boldsymbol{\eta} \in R^k : -\infty < A(\boldsymbol{\eta}) < \infty\}.$$

Example 1.6.5. *($\mathcal{N}(\mu,\sigma^2)$ continued).* In this example, $k=2$, $\mathbf{T}^T(x)=(x,x^2)=(T_1(x),T_2(x))$, $\eta_1=\mu/\sigma^2$, $\eta_2=-1/2\sigma^2$, $A(\boldsymbol{\eta})=\frac{1}{2}[(-\eta_1^2/2\eta_2)+\log(\pi/-\eta_2)]$, $h(x)=1$ and $\mathcal{E}=R\times R^-=\{(\eta_1,\eta_2):\eta_1\in R,\eta_2<0\}$.

Example 1.6.6. *Linear Regression.* Suppose as in Examples 1.1.4 and 1.5.5 that $Y_1,\ldots,Y_n$ are independent, $Y_i\sim\mathcal{N}(\mu_i,\sigma^2)$, with $\mu_i=\beta_1+\beta_2 z_i$, $i=1,\ldots,n$. From Example 1.5.5, the density of $\mathbf{Y}=(Y_1,\ldots,Y_n)^T$ can be put in canonical form with $k=3$, $\mathbf{T}(\mathbf{Y})=(\Sigma Y_i,\Sigma Y_i^2,\Sigma z_iY_i)^T$, $\eta_1=\beta_1/\sigma^2$, $\eta_2=\beta_2/\sigma^2$, $\eta_3=-1/2\sigma^2$,

$$A(\boldsymbol{\eta})=\frac{-n}{4\eta_3}[\eta_1^2+\widehat{m}_2\eta_2^2+\bar{z}\eta_1\eta_2+2\log(\pi/-\eta_3)],$$

and $\mathcal{E}=\{(\eta_1,\eta_2,\eta_3):\eta_1\in R,\eta_2\in R,\eta_3<0\}$, where $\widehat{m}_2=n^{-1}\Sigma z_i^2$.

Example 1.6.7. *Multinomial Trials.* We observe the outcomes of n independent trials where each trial can end up in one of k possible categories. We write the outcome vector as $\mathbf{X}=(X_1,\ldots,X_n)^T$ where the X_i are i.i.d. as X and the sample space of each X_i is the k categories $\{1,2,\ldots,k\}$. Let $T_j(\mathbf{x})=\sum_{i=1}^n 1[X_i=j]$, and $\lambda_j=P(X_i=j)$. Then $p(\mathbf{x},\boldsymbol{\lambda})=\prod_{j=1}^k\lambda_j^{T_j(\mathbf{x})}$, $\boldsymbol{\lambda}\in\Lambda$, where Λ is the simplex $\{\boldsymbol{\lambda}\in R^k:0<\lambda_j<1,j=1,\ldots,k,\sum_{j=1}^k\lambda_j=1\}$. It will often be more convenient to work with unrestricted parameters. In this example, we can achieve this by the reparametrization

$$\lambda_j=e^{\alpha_j}/\sum_{j=1}^k e^{\alpha_j},\ j=1,\ldots,k,\boldsymbol{\alpha}\in R^k.$$

Now we can write the likelihood as

$$q_0(\mathbf{x},\boldsymbol{\alpha})=\exp\{\sum_{j=1}^k\alpha_jT_j(\mathbf{x})-n\log\sum_{j=1}^k\exp(\alpha_j)\}.$$

This is a k-parameter canonical exponential family generated by $T_1,\ldots,T_k$ and $h(\mathbf{x})=\prod_{i=1}^n 1[x_i\in\{1,\ldots,k\}]$ with canonical parameter $\boldsymbol{\alpha}$ and $\mathcal{E}=R^k$. However $\boldsymbol{\alpha}$ is not identifiable because $q_0(\mathbf{x},\boldsymbol{\alpha}+c\mathbf{1})=q_0(\mathbf{x},\boldsymbol{\alpha})$ for $\mathbf{1}=(1,\ldots,1)^T$ and all c. This can be remedied by considering

$$\mathbf{T}_{(k-1)}(\mathbf{x})\equiv(T_1(\mathbf{x}),\ldots,T_{k-1}(\mathbf{x}))^T,$$

$\eta_j=\log(\lambda_j/\lambda_k)=\alpha_j-\alpha_k$, $1\le j\le k-1$, and rewriting

$$q(\mathbf{x},\boldsymbol{\eta})=\exp\{\mathbf{T}_{(k-1)}^T(\mathbf{x})\boldsymbol{\eta}-n\log(1+\sum_{j=1}^{k-1}e^{\eta_j})\}$$

where

$$\lambda_j=\frac{e^{\eta_j}}{1+\sum_{j=1}^{k-1}e^{\eta_j}}=\frac{e^{\alpha_j}}{\sum_{j=1}^k e^{\alpha_j}},\ j=1,\ldots,k-1.$$

Note that $q(\mathbf{x}, \boldsymbol{\eta})$ is a $k-1$ parameter canonical exponential family generated by $\mathbf{T}_{(k-1)}$ and $h(\mathbf{x}) = \prod_{i=1}^{n} 1[x_i \in \{1, \ldots, k\}]$ with canonical parameter $\boldsymbol{\eta}$ and $\mathcal{E} = R^{k-1}$. Moreover, the parameters $\eta_j = \log(P_{\boldsymbol{\eta}}[X = j]/P_{\boldsymbol{\eta}}[X = k])$, $1 \leq j \leq k-1$, are identifiable. Note that the model for $\mathbf{X}$ is unchanged. □

1.6.3 Building Exponential Families

Submodels

A *submodel* of a k-parameter canonical exponential family $\{q(\mathbf{x}, \boldsymbol{\eta}); \boldsymbol{\eta} \in \mathcal{E} \subset R^k\}$ is an exponential family defined by

$$p(x, \boldsymbol{\theta}) = q(x, \boldsymbol{\eta}(\boldsymbol{\theta})) \tag{1.6.12}$$

where $\boldsymbol{\theta} \in \Theta \subset R^l$, $l \leq k$, and $\boldsymbol{\eta}$ is a map from Θ to a subset of R^k. Thus, if X is discrete taking on k values as in Example 1.6.7 and $\mathbf{X} = (X_1, \ldots, X_n)^T$ where the X_i are i.i.d. as X, then *all* models for $\mathbf{X}$ are exponential families because they are submodels of the multinomial trials model.

Affine transformations

If $\mathcal{P}$ is the canonical family generated by $\mathbf{T}_{k\times 1}$ and h and $\mathbf{M}$ is the affine transformation from R^k to R^l defined by

$$\mathbf{M}(\mathbf{T}) = M_{\ell\times k}\mathbf{T} + \mathbf{b}_{\ell\times 1},$$

it is easy to see that the family generated by $\mathbf{M}(\mathbf{T}(X))$ and h is the subfamily of $\mathcal{P}$ corresponding to

$$\Theta = [\boldsymbol{\eta}^{-1}](\mathcal{E}) \subset R^{\ell}$$

and

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = M^T\boldsymbol{\theta}.$$

Similarly, if $\Theta \subset R^{\ell}$ and $\boldsymbol{\eta}(\boldsymbol{\theta}) = B_{k\times\ell}\boldsymbol{\theta} \subset R^k$, then the resulting submodel of $\mathcal{P}$ above is a submodel of the exponential family generated by $B^T\mathbf{T}(X)$ and h. See Problem 1.6.17 for details. Here is an example of affine transformations of $\boldsymbol{\theta}$ and $\mathbf{T}$.

Example 1.6.8. *Logistic Regression.* Let Y_i be independent binomial, $\mathcal{B}(n_i, \lambda_i)$, $1 \leq i \leq n$. If the λ_i are unrestricted, $0 < \lambda_i < 1$, $1 \leq i \leq n$, this, from Example 1.6.2, is an n-parameter canonical exponential family with $\mathcal{Y}_i \equiv$ integers from 0 to n_i generated by $\mathbf{T}(Y_1, \ldots, Y_n) = \mathbf{Y}, h(\mathbf{y}) = \prod_{i=1}^{n} \binom{n_i}{y_i} 1(0 \leq y_i \leq n_i)$. Here $\eta_i = \log \frac{\lambda_i}{1-\lambda_i}$, $A(\boldsymbol{\eta}) = \sum_{i=1}^{n} n_i \log(1 + e^{\eta_i})$. However, let $x_1 < \ldots < x_n$ be specified levels and

$$\eta_i(\boldsymbol{\theta}) = \theta_1 + \theta_2 x_i, \quad 1 \leq i \leq n, \quad \boldsymbol{\theta} = (\theta_1, \theta_2)^T \in R^2. \tag{1.6.13}$$

This is a linear transformation $\boldsymbol{\eta}(\boldsymbol{\theta}) = B_{n\times 2}\boldsymbol{\theta}$ corresponding to $B_{n\times 2} = (\mathbf{1}, \mathbf{x})$, where $\mathbf{1}$ is $(1, \ldots, 1)^T$, $\mathbf{x} = (x_1, \ldots, x_n)^T$. Set $M = B^T$, then this is the two-parameter canonical exponential family generated by $M\mathbf{Y} = (\sum_{i=1}^n Y_i, \sum_{i=1}^n x_i Y_i)^T$ and h with

$$A(\theta_1, \theta_2) = \sum_{i=1}^n n_i \log(1 + \exp(\theta_1 + \theta_2 x_i)).$$

This model is sometimes applied in experiments to determine the toxicity of a substance. The Y_i represent the number of animals dying out of n_i when exposed to level x_i of the substance.It is assumed that each animal has a random toxicity threshold X such that death results if and only if a substance level on or above X is applied. Assume also:

(a) No interaction between animals (independence) in relation to drug effects

(b) The distribution of X in the animal population is *logistic*; that is,

$$P[X \leq x] = [1 + \exp\{-(\theta_1 + \theta_2 x)\}]^{-1}, \tag{1.6.14}$$

$\theta_1 \in R$, $\theta_2 > 0$. Then (and only then),

$$\log(P[X \leq x]/(1 - P[X \leq x])) = \theta_1 + \theta_2 x$$

and (1.6.13) holds. □

Curved exponential families

Exponential families (1.6.12) with the range of $\boldsymbol{\eta}(\boldsymbol{\theta})$ restricted to a subset of dimension l with $l \leq k - 1$, are called *curved exponential families* provided they do not form a canonical exponential family in the $\boldsymbol{\theta}$ parametrization.

Example 1.6.9. *Gaussian with Fixed Signal-to-Noise Ratio.* In the normal case with $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, suppose the ratio $|\mu|/\sigma$, which is called the *coefficient of variation* or *signal-to-noise ratio*, is a known constant $\lambda_0 > 0$. Then, with $\theta = \mu$, we can write

$$p(\mathbf{x}, \theta) = \exp\left\{\lambda_0^2\theta^{-1}T_1 - \frac{1}{2}\lambda_0^2\theta^{-2}T_2 - \frac{1}{2}n[\lambda_0^2 + \log(2\pi\lambda_0^2\theta^2)]\right\}$$

where $T_1 = \sum_{i=1}^n x_i$, $T_2 = \sum_{i=1}^n x_i^2$, $\eta_1(\theta) = \lambda_0^2\theta^{-1}$ and $\eta_2(\theta) = -\frac{1}{2}\lambda_0^2\theta^{-2}$. This is a curved exponential family with $l = 1$. □

In Example 1.6.8, the $\boldsymbol{\theta}$ parametrization has dimension 2, which is less than $k = n$ when $n > 3$. However, $p(x, \boldsymbol{\theta})$ in the $\boldsymbol{\theta}$ parametrization is a canonical exponential family, so it is not a curved family.

Example 1.6.10. *Location-Scale Regression.* Suppose that $Y_1, \ldots, Y_n$ are independent, $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. If each μ_i ranges over R and each σ_i^2 ranges over $(0, \infty)$, this is by Example 1.6.5 a $2n$-parameter canonical exponential family model with $\eta_i = \mu_i/\sigma_i^2$, and $\eta_{n+i} = -1/2\sigma_i^2$, $i = 1, \ldots, n$, generated by

$$T(\mathbf{Y}) = (Y_1, \ldots, Y_n, Y_1^2, \ldots, Y_n^2)^T$$

and $h(\mathbf{Y}) = 1$. Next suppose that (μ_i, σ_i^2) depend on the value z_i of some covariate, say,

$$\mu_i = \theta_1 + \theta_2 z_i,\ \sigma_i^2 = \theta_3(\theta_1 + \theta_2 z_i)^2,\ z_1 < \cdots < z_n$$

for unknown parameters $\theta_1 \in R$, $\theta_2 \in R$, $\theta_3 > 0$ (e.g., Bickel, 1978; Carroll and Ruppert, 1988, Sections 2.1–2.5; and Snedecor and Cochran, 1989, Section 15.10). For $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_2)$, the map $\boldsymbol{\eta}(\boldsymbol{\theta})$ is

$$\eta_i(\boldsymbol{\theta}) = \theta_3^{-1}(\theta_1 + \theta_2 z_i)^{-1},\ \eta_{n+i}(\boldsymbol{\theta}) = \frac{1}{2}\theta_3^{-1}(\theta_1 + \theta_2 z_i)^{-2},\ i = 1, \ldots, n.$$

Because $\sum_{i=1}^n \eta_i(\boldsymbol{\theta})Y_i + \sum_{i=1}^n \eta_{n+i}(\boldsymbol{\theta})Y_i^2$ cannot be written in the form $\sum_{j=1}^n \eta_j^*(\boldsymbol{\theta})T_j^*(\mathbf{Y})$ for some $\eta_j^*(\boldsymbol{\theta})$, $T_j^*(\mathbf{Y})$, then $p(\mathbf{y}, \boldsymbol{\theta}) = q(\mathbf{y}, \boldsymbol{\eta}(\boldsymbol{\theta}))$ as defined in (6.1.12) is not an exponential family model, but a curved exponential family model with $l = 3$. □

Models in which the variance $\mathrm{Var}(Y_i)$ depends on i are called *heteroscedastic* whereas models in which $\mathrm{Var}(Y_i)$ does not depend on i are called *homoscedastic*. Thus, Examples 1.6.10 and 1.6.6 are heteroscedastic and homoscedastic models, respectively.

We return to curved exponential family models in Section 2.3.

Supermodels

We have already noted that the exponential family structure is preserved under i.i.d. sampling. Even more is true. Let Y_j, $1 \le j \le n$, be independent, $Y_j \in \mathcal{Y}_j \subset R^q$, with an exponential family density

$$q_j(y_j, \boldsymbol{\theta}) = \exp\{\mathbf{T}_j^T(y_j)\boldsymbol{\eta}(\boldsymbol{\theta}) - B_j(\boldsymbol{\theta})\}h_j(y_j),\ \boldsymbol{\theta} \in \Theta \subset R^k.$$

Then $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)^T$ is modeled by the exponential family generated by $\mathbf{T}(\mathbf{Y}) = \sum_{j=1}^n \mathbf{T}_j(Y_j)$ and $\prod_{j=1}^n h_j(y_j)$, with parameter $\boldsymbol{\eta}(\boldsymbol{\theta})$, and $B(\boldsymbol{\theta}) = \sum_{j=1}^n B_j(\boldsymbol{\theta})$.

In Example 1.6.8 note that (1.6.13) exhibits Y_j as being distributed according to a two-parameter family generated by $T_j(Y_j) = (Y_j, x_j Y_j)$ and we can apply the supermodel approach to reach the same conclusion as before.

1.6.4 Properties of Exponential Families

Theorem 1.6.1 generalizes directly to k-parameter families as does its continuous analogue. We extend the statement of Theorem 1.6.2.

Recall from Section B.5 that for any random vector $\mathbf{T}_{k\times 1}$, we define

$$M(\mathbf{s}) \equiv Ee^{\mathbf{s}^T\mathbf{T}}$$

as the moment-generating function, and

$$E(\mathbf{T}) \equiv (E(T_1), \ldots, E(T_k))^T$$

$$\mathrm{Var}(\mathbf{T}) = \|\mathrm{Cov}(T_a, T_b)\|_{k\times k}.$$

Theorem 1.6.3. *Let $\mathcal{P}$ be a canonical k-parameter exponential family generated by $(\mathbf{T}, h)$ with corresponding natural parameter space $\mathcal{E}$ and function $A(\boldsymbol{\eta})$. Then*

(a) *$\mathcal{E}$ is convex*

(b) *$A : \mathcal{E} \to R$ is convex*

(c) *If $\mathcal{E}$ has nonempty interior in R^k and $\boldsymbol{\eta}_0 \in \mathcal{E}$, then $\mathbf{T}(X)$ has under $\boldsymbol{\eta}_0$ a moment-generating function M given by*

$$M(\mathbf{s}) = \exp\{A(\boldsymbol{\eta}_0 + \mathbf{s}) - A(\boldsymbol{\eta}_0)\}$$

valid for all $\mathbf{s}$ such that $\boldsymbol{\eta}_0 + \mathbf{s} \in \mathcal{E}$. Since $\boldsymbol{\eta}_0$ is an interior point this set of $\mathbf{s}$ includes a ball about $\mathbf{0}$.

Corollary 1.6.1. *Under the conditions of Theorem* 1.6.3

$$E_{\boldsymbol{\eta}_0}\mathbf{T}(X) = \dot{A}(\boldsymbol{\eta}_0)$$

$$\mathrm{Var}_{\boldsymbol{\eta}_0}\mathbf{T}(X) = \ddot{A}(\boldsymbol{\eta}_0)$$

where $\dot{A}(\boldsymbol{\eta}_0) = (\frac{\partial A}{\partial \eta_1}(\boldsymbol{\eta}_0), \ldots, \frac{\partial A}{\partial \eta_k}(\boldsymbol{\eta}_0))^T$, $\ddot{A}(\boldsymbol{\eta}_0) = \|\frac{\partial^2 A}{\partial \eta_a \partial \eta_b}(\boldsymbol{\eta}_0)\|$.

The corollary follows immediately from Theorem B.5.1 and Theorem 1.6.3(c).

Proof of Theorem 1.6.3. We prove (b) first. Suppose $\boldsymbol{\eta}_0, \boldsymbol{\eta}_1 \in \mathcal{E}$ and $0 \le \alpha \le 1$. By the Hölder inequality (B.9.4), for any $u(x), v(x), h(x) \ge 0$, $r, s > 0$ with $\frac{1}{r} + \frac{1}{s} = 1$,

$$\int u(x)v(x)h(x)dx \le \left(\int u^r(x)h(x)dx\right)^{\frac{1}{r}}\left(\int v^s(x)h(x)dx\right)^{\frac{1}{s}}.$$

Substitute $\frac{1}{r} = \alpha$, $\frac{1}{s} = 1 - \alpha$, $u(x) = \exp(\alpha\boldsymbol{\eta}_1^T\mathbf{T}(x))$, $v(x) = \exp((1-\alpha)\boldsymbol{\eta}_2^T\mathbf{T}(x))$ and take logs of both sides to obtain, (with ∞ permitted on either side),

$$A(\alpha\boldsymbol{\eta}_1 + (1-\alpha)\boldsymbol{\eta}_2) \le \alpha A(\boldsymbol{\eta}_1) + (1-\alpha)A(\boldsymbol{\eta}_2) \tag{1.6.15}$$

which is (b). If $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathcal{E}$ the right-hand side of (1.6.15) is finite. Because

$$\int \exp(\boldsymbol{\eta}^T\mathbf{T}(x))h(x)dx > 0$$

for all $\boldsymbol{\eta}$ we conclude from (1.6.15) that $\alpha\boldsymbol{\eta}_1 + (1-\alpha)\boldsymbol{\eta}_2 \in \mathcal{E}$ and (a) follows. Finally (c) is proved in exactly the same way as Theorem 1.6.2. □

The formulae of Corollary 1.6.1 give a classical result in Example 1.6.6.

Example 1.6.7. *(continued).* Here, using the α parametrization,

$$A(\alpha) = n \log(\sum_{j=1}^{k} e^{\alpha_j})$$

and

$$E_{\boldsymbol{\lambda}}(T_j(\mathbf{X})) = P_{\boldsymbol{\lambda}}[X = j] \equiv \lambda_j = e^{\alpha_j} / \sum_{\ell=1}^{k} e^{\alpha_\ell}$$

$$\begin{aligned} \mathrm{Cov}_{\boldsymbol{\lambda}}(T_i, T_j) &= \frac{\partial^2 A}{\partial \alpha_i \partial \alpha_j}(\alpha) = -n \frac{e^{\alpha_j} e^{\alpha_i}}{(\sum_{\ell=1}^{k} e^{\alpha_\ell})^2} = -n\lambda_i\lambda_j, \quad i \neq j \\ \mathrm{Var}_{\boldsymbol{\lambda}}(T_i) &= \frac{\partial^2 A}{\partial \alpha_i^2}(\alpha) = n\lambda_i(1 - \lambda_i). \end{aligned}$$

□

The rank of an exponential family

Evidently every k-parameter exponential family is also k'-dimensional with $k' > k$. However, there is a minimal dimension.

An exponential family is of *rank* k iff the generating statistic $\mathbf{T}$ is k-dimensional and 1, $T_1(X), \ldots, T_k(X)$ are linearly independent with positive probability. Formally, $P_{\boldsymbol{\eta}}[\sum_{j=1}^{k} a_j T_j(X) = a_{k+1}] < 1$ unless all a_j are 0.

Note that $P_{\boldsymbol{\theta}}(A) = 0$ or $P_{\boldsymbol{\theta}}(A) < 1$ for some $\boldsymbol{\theta}$ iff the corresponding statement holds for all $\boldsymbol{\theta}$ because $0 < \frac{p(x,\boldsymbol{\theta}_1)}{p(x,\boldsymbol{\theta}_2)} < \infty$ for all x, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ such that $h(x) > 0$.

Going back to Example 1.6.7 we can see that the multinomial family is of rank at most $k - 1$. It is intuitively clear that $k - 1$ is in fact its rank and this is seen in Theorem 1.6.4 that follows. Similarly, in Example 1.6.8, if $n = 1$, and $\eta_1(\boldsymbol{\theta}) = \theta_1 + \theta_2 x_1$ we are writing the one-parameter binomial family corresponding to Y_1 as a two-parameter family with generating statistic $(Y_1, x_1 Y_1)$. But the rank of the family is 1 and θ_1 and θ_2 are not identifiable. However, if we consider $\mathbf{Y}$ with $n \geq 2$ and $x_1 < x_n$ the family as we have seen remains of rank ≤ 2 and is in fact of rank 2. Our discussion suggests a link between rank and identifiability of the $\boldsymbol{\eta}$ parameterization. We establish the connection and other fundamental relationships in Theorem 1.6.4.

Theorem 1.6.4. *Suppose $\mathcal{P} = \{q(x, \boldsymbol{\eta});\ \boldsymbol{\eta} \in \mathcal{E}\}$ is a canonical exponential family generated by $(\mathbf{T}_{k\times 1}, h)$ with natural parameter space $\mathcal{E}$ such that $\mathcal{E}$ is open. Then the following are equivalent.*

(i) *$\mathcal{P}$ is of rank k.*

(ii) *$\boldsymbol{\eta}$ is a parameter (identifiable).*

(iii) *$\mathrm{Var}_{\boldsymbol{\eta}}(\mathbf{T})$ is positive definite.*

(iv) $\eta \to \dot{A}(\eta)$ *is* 1-1 *on* $\mathcal{E}$.

(v) A *is strictly convex on* $\mathcal{E}$.

Note that, by Theorem 1.6.3, *because* $\mathcal{E}$ *is open,* $\dot{A}$ *is defined on all of* $\mathcal{E}$.

Proof. We give a detailed proof for $k = 1$. The proof for $k > 1$ is then sketched with details left to a problem. Let $\sim(\cdot)$ denote "$(\cdot)$ is false." Then

$\sim$(i) $\Leftrightarrow P_\eta[a_1 T = a_2] = 1$ for $a_1 \neq 0$. This is equivalent to $\mathrm{Var}_\eta(T) = 0 \Leftrightarrow \sim$ (iii)

$\sim$(ii) $\Leftrightarrow$ There exist $\eta_1 \neq \eta_2$ such that $P_{\eta_1} = P_{\eta_2}$.

Equivalently

$$\exp\{\eta_1 T(x) - A(\eta_1)\}h(x) = \exp\{\eta_2 T(x) - A(\eta_2)\}h(x).$$

Taking logs we obtain $(\eta_1 - \eta_2)T(X) = A(\eta_2) - A(\eta_1)$ with probability $1 \equiv \sim$(i). We, thus, have (i) $\equiv$ (ii) $\equiv$ (iii). Now (iii) $\Rightarrow A''(\eta) > 0$ by Theorem 1.6.2 and, hence, $A'(\eta)$ is strictly monotone increasing and 1-1. Conversely, $A''(\eta_0) = 0$ for some η_0 implies that $T \equiv c$, with probability 1, for all η, by our remarks in the discussion of rank, which implies that $A''(\eta) = 0$ for all η and, hence, A' is constant. Thus, (iii) $\equiv$ (iv) and the same discussion shows that (iii) $\equiv$ (v).

Proof of the general case sketched

I. $\sim$ (i) $\equiv \sim$ (iii)

$\sim$ (i) $\equiv P_{\boldsymbol{\eta}}[\mathbf{a}^T\mathbf{T} = c] = 1$ for some $\mathbf{a} \neq 0$, all $\boldsymbol{\eta}$
$\sim$ (iii) $\equiv \mathbf{a}^T\mathrm{Var}_{\boldsymbol{\eta}}(\mathbf{T})\mathbf{a} = \mathrm{Var}_{\boldsymbol{\eta}}(\mathbf{a}^T\mathbf{T}) = 0$ for some $\mathbf{a} \neq 0$, all $\boldsymbol{\eta} \equiv (\sim i)$

II. $\sim$ (ii) $\equiv \sim$ (i)

$\sim$ (ii) $\equiv P_{\boldsymbol{\eta}_1} = P_{\boldsymbol{\eta}_0}$ some $\boldsymbol{\eta}_1 \neq \boldsymbol{\eta}_0$. Let

$$\mathcal{Q} = \{P_{\boldsymbol{\eta}_0 + c(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_0)} : \boldsymbol{\eta}_0 + c(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_0) \in \mathcal{E}\}.$$

$\mathcal{Q}$ is the exponential family (one-parameter) generated by $(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_0)^T\mathbf{T}$. Apply the case $k = 1$ to $\mathcal{Q}$ to get $\sim$ (ii) $\equiv \sim$ (i).

III. (iv) $\equiv$ (v) $\equiv$ (iii)

Properties (iv) and (v) are equivalent to the statements holding for every $\mathcal{Q}$ defined as previously for arbitrary $\boldsymbol{\eta}_0, \boldsymbol{\eta}_1$. □

Corollary 1.6.2. *Suppose that the conditions of Theorem* 1.6.4 *hold and* $\mathcal{P}$ *is of rank* k. *Then*

(a) $\mathcal{P}$ *may be uniquely parametrized by* $\boldsymbol{\mu}(\boldsymbol{\eta}) \equiv E_{\boldsymbol{\eta}}\mathbf{T}(X)$ *where* $\boldsymbol{\mu}$ *ranges over* $\dot{A}(\mathcal{E})$,
(b) $\log q(x, \boldsymbol{\eta})$ *is a strictly concave function of* $\boldsymbol{\eta}$ *on* $\mathcal{E}$.

Proof. This is just a restatement of (iv) and (v) of the theorem. □

The relation in (a) is sometimes evident and the μ parametrization is close to the initial parametrization of classical $\mathcal{P}$. Thus, the $\mathcal{B}(n,\theta)$ family is parametrized by $E(X)$, where X is the Bernoulli trial, the $\mathcal{N}(\mu,\sigma_0^2)$ family by $E(X)$. For $\{\mathcal{N}(\mu,\sigma^2)\}$, $E(X,X^2) = (\mu,\sigma^2+\mu^2)$, which is obviously a 1-1 function of (μ,σ^2). However, the relation in (a) may be far from obvious (see Problem 1.6.21). The corollary will prove very important in estimation theory. See Section 2.3. We close the present discussion of exponential families with the following example.

Example 1.6.11. *The p Variate Gaussian Family.* An important exponential family is based on the multivariate Gaussian distributions of Section B.6. Recall that $\mathbf{Y}_{p\times 1}$ has a p variate Gaussian distribution, $\mathcal{N}_p(\boldsymbol{\mu},\Sigma)$, with mean $\boldsymbol{\mu}_{p\times 1}$ and positive definite variance covariance matrix $\sum_{p\times p}$, iff its density is

$$f(\mathbf{Y},\boldsymbol{\mu},\Sigma) = |\det(\Sigma)|^{-1/2}\pi^{-p/2}\exp\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{Y}-\boldsymbol{\mu})\}. \tag{1.6.16}$$

Rewriting the exponent we obtain

$$\begin{aligned}\log f(\mathbf{Y},\boldsymbol{\mu},\Sigma) &= -\frac{1}{2}\mathbf{Y}^T\Sigma^{-1}\mathbf{Y} + (\Sigma^{-1}\boldsymbol{\mu})^T\mathbf{Y} \\ &- \frac{1}{2}(\log|\det(\Sigma)| + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}) - \frac{p}{2}\log\pi.\end{aligned} \tag{1.6.17}$$

The first two terms on the right in (1.6.17) can be rewritten

$$-(\sum_{1\le i<j\le p}\sigma^{ij}Y_iY_j + \frac{1}{2}\sum_{i=1}^p\sigma^{ii}Y_i^2) + \sum_{i=1}^p(\sum_{j=1}^p\sigma^{ij}\mu_j)Y_i$$

where $\Sigma^{-1} \equiv \|\sigma^{ij}\|$, revealing that this is a $k = p(p+3)/2$ parameter exponential family with statistics $(Y_1,\ldots,Y_p,\{Y_iY_j\}_{1\le i\le j\le p})$, $h(\mathbf{Y}) \equiv 1$, $\boldsymbol{\theta} = (\boldsymbol{\mu},\Sigma)$, $B(\boldsymbol{\theta}) = \frac{1}{2}(\log|\det(\Sigma)| + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu})$. By our supermodel discussion, if $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ are iid $\mathcal{N}_p(\boldsymbol{\mu},\Sigma)$, then $\mathbf{X} \equiv (\mathbf{Y}_1,\ldots,\mathbf{Y}_n)^T$ follows the $k = p(p+3)/2$ parameter exponential family with $\mathbf{T} = (\Sigma_i\mathbf{Y}_i, \Sigma_i\mathbf{Y}_i\mathbf{Y}_i^T)$, where we identify the second element of $\mathbf{T}$, which is a $p\times p$ symmetric matrix, with its distinct $p(p+1)/2$ entries. It may be shown (Problem 1.6.29) that $\mathbf{T}$ (and $h\equiv 1$) generate this family and that the rank of the family is indeed $p(p+3)/2$, generalizing Example 1.6.5, and that $\mathcal{E}$ is open, so that Theorem 1.6.4 applies. □

1.6.5 Conjugate Families of Prior Distributions

In Section 1.2 we considered beta prior distributions for the probability of success in n Bernoulli trials. This is a special case of *conjugate families* of priors, families to which the posterior after sampling also belongs.

Suppose $X_1,\ldots,X_n$ is a sample from the k-parameter exponential family (1.6.10), and, as we always do in the Bayesian context, write $p(\mathbf{x}\mid\theta)$ for $p(\mathbf{x},\theta)$. Then

$$p(\mathbf{x}|\boldsymbol{\theta}) = [\prod_{i=1}^n h(x_i)]\exp\{\sum_{j=1}^k\eta_j(\boldsymbol{\theta})\sum_{i=1}^n T_j(x_i) - nB(\boldsymbol{\theta})\}. \tag{1.6.18}$$

where $\boldsymbol{\theta} \in \Theta$, which is k-dimensional. A conjugate exponential family is obtained from (1.6.18) by letting n and $t_j = \sum_{i=1}^n T_j(x_i)$, $j = 1, \ldots, k$, be "parameters" and treating $\boldsymbol{\theta}$ as the variable of interest. That is, let $\mathbf{t} = (t_1, \ldots, t_{k+1})^T$ and

$$\begin{aligned} \omega(\mathbf{t}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\{\sum_{j=1}^{k} t_j \eta_j(\boldsymbol{\theta}) - t_{k+1} B(\boldsymbol{\theta})\} d\theta_1 \cdots d\theta_k \\ \Omega &= \{(t_1, \ldots, t_{k+1}) : 0 < \omega(t_1, \ldots, t_{k+1}) < \infty\} \end{aligned} \tag{1.6.19}$$

with integrals replaced by sums in the discrete case. We assume that Ω is nonempty (see Problem 1.6.36), then

Proposition 1.6.1. *The $(k+1)$-parameter exponential family given by*

$$\pi_{\mathbf{t}}(\boldsymbol{\theta}) = \exp\{\sum_{j=1}^{k} \eta_j(\boldsymbol{\theta}) t_j - t_{k+1} B(\boldsymbol{\theta}) - \log \omega(\mathbf{t})\} \tag{1.6.20}$$

where $\mathbf{t} = (t_1, \ldots, t_{k+1}) \in \Omega$, is a conjugate prior to $p(\mathbf{x}|\boldsymbol{\theta})$ given by (1.6.18).

Proof. If $p(\mathbf{x}|\boldsymbol{\theta})$ is given by (1.6.18) and π by (1.6.20), then

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{x}) &\propto p(\mathbf{x}|\boldsymbol{\theta}) \pi_{\mathbf{t}}(\boldsymbol{\theta}) \propto \exp\{\sum_{j=1}^{k} \eta_j(\boldsymbol{\theta})(\sum_{i=1}^{n} T_j(x_i) + t_j) - (t_{k+1} + n) B(\boldsymbol{\theta})\} \\ &\propto \pi_{\mathbf{s}}(\boldsymbol{\theta}), \end{aligned} \tag{1.6.21}$$

where

$$\mathbf{s} = (s_1, \ldots, s_{k+1})^T = \left(t_1 + \sum_{i=1}^{n} T_1(x_i), \ldots, t_k + \sum_{i=1}^{n} T_k(x_i), t_{k+1} + n\right)^T$$

and $\propto$ indicates that the two sides are proportional functions of $\boldsymbol{\theta}$. Because two probability densities that are proportional must be equal, $\pi(\boldsymbol{\theta}|\mathbf{x})$ is the member of the exponential family (1.6.20) given by the last expression in (1.6.21) and our assertion follows. □

Remark 1.6.1. Note that (1.6.21) is an updating formula in the sense that as data $x_1, \ldots, x_n$ become available, the parameter $\mathbf{t}$ of the prior distribution is updated to $\mathbf{s} = (\mathbf{t} + \mathbf{a})$, where $\mathbf{a} = (\sum_{i=1}^n T_1(x_i), \ldots, \sum_{i=1}^n T_k(x_i), n)^T$. □

It is easy to check that the beta distributions are obtained as conjugate to the binomial in this way.

Example 1.6.12. Suppose $X_1, \ldots, X_n$ is a $\mathcal{N}(\theta, \sigma_0^2)$ sample, where σ_0^2 is known and θ is unknown. To choose a prior distribution for θ, we consider the conjugate family of the model defined by (1.6.20). For $n = 1$

$$p(x|\theta) \propto \exp\{\frac{\theta x}{\sigma_0^2} - \frac{\theta^2}{2\sigma_0^2}\}. \tag{1.6.22}$$

This is a one-parameter exponential family with

$$T_1(x) = x, \eta_1(\theta) = \frac{\theta}{\sigma_0^2}, B(\theta) = \frac{\theta^2}{2\sigma_0^2}.$$

The conjugate two-parameter exponential family given by (1.6.20) has density

$$\pi_t(\theta) = \exp\{\frac{\theta}{\sigma_0^2}t_1 - \frac{\theta^2}{2\sigma_0^2}t_2 - \log\omega(t_1, t_2)\}. \tag{1.6.23}$$

Upon completing the square, we obtain

$$\pi_t(\theta) \propto \exp\{-\frac{t_2}{2\sigma_0^2}(\theta - \frac{t_1}{t_2})^2\}. \tag{1.6.24}$$

Thus, $\pi_t(\theta)$ is defined only for $t_2 > 0$ and all t_1 and is the $\mathcal{N}(t_1/t_2, \sigma_0^2/t_2)$ density. Our conjugate family, therefore, consists of all $\mathcal{N}(\eta_0, \tau_0^2)$ distributions where η_0 varies freely and τ_0^2 is positive.

If we start with a $\mathcal{N}(\eta_0, \tau_0^2)$ prior density, we must have in the (t_1, t_2) parametrization

$$t_2 = \frac{\sigma_0^2}{\tau_0^2}, \quad t_1 = \frac{\eta_0\sigma_0^2}{\tau_0^2}. \tag{1.6.25}$$

By (1.6.21), if we observe $\Sigma X_i = s$, the posterior has a density (1.6.23) with

$$t_2(n) = \frac{\sigma_0^2}{\tau_0^2} + n, \quad t_1(s) = \frac{\eta_0\sigma_0^2}{\tau_0^2} + s.$$

Using (1.6.24), we find that $\pi(\theta|\mathbf{x})$ is a normal density with mean

$$\mu(s, n) = \frac{t_1(s)}{t_2(n)} = (\frac{\sigma_0^2}{\tau_0^2} + n)^{-1}[s + \frac{\eta_0\sigma_0^2}{\tau_0^2}] \tag{1.6.26}$$

and variance

$$\tau_0^2(n) = \frac{\sigma_0^2}{t_2(n)} = (\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2})^{-1}. \tag{1.6.27}$$

Note that we can rewrite (1.6.26) intuitively as

$$\mu(s, n) = w_1\bar{x} + w_2\eta_0 \tag{1.6.28}$$

where $w_1 = n\tau_0^2(n)/\sigma_0^2$, $w_2 = \tau_0^2(n)/\tau_0^2$ so that $w_2 = 1 - w_1$. □

These formulae can be generalized to the case $\mathbf{X}_i$ i.i.d. $\mathcal{N}_p(\boldsymbol{\theta}, \Sigma_0)$, $1 \leq i \leq n$, Σ_0 known, $\boldsymbol{\theta} \sim \mathcal{N}_p(\boldsymbol{\eta}_0, \tau_0^2\mathbf{I})$ where $\boldsymbol{\eta}_0$ varies over R^p, τ_0^2 is scalar with $\tau_0 > 0$ and $\mathbf{I}$ is the $p \times p$ identity matrix (Problem 1.6.37). Moreover, it can be shown (Problem 1.6.30) that the $\mathcal{N}_p(\boldsymbol{\lambda}, \Gamma)$ family with $\boldsymbol{\lambda} \in R^p$ and Γ symmetric positive definite is a conjugate family

to $\mathcal{N}_p(\boldsymbol{\theta}, \Sigma_0)$, but a richer one than we've defined in (1.6.20) except for $p = 1$ because $\mathcal{N}_p(\boldsymbol{\lambda}, \Gamma)$ is a $p(p+3)/2$ rather than a $p+1$ parameter family. In fact, the conditions of Proposition 1.6.1 are often too restrictive. In the one-dimensional Gaussian case the members of the Gaussian conjugate family are unimodal and symmetric and have the same shape. It is easy to see that one can construct conjugate priors for which one gets reasonable formulae for the parameters indexing the model and yet have as great a richness of the shape variable as one wishes by considering finite mixtures of members of the family defined in (1.6.20). See Problems 1.6.31 and 1.6.32.

(1.6.23)

Discussion

Note that the uniform $\mathcal{U}(\{1, 2, \ldots, \theta\})$ model of Example 1.5.3 is *not* covered by this theory. The natural sufficient statistic $\max(X_1, \ldots, X_n)$, which is one-dimensional whatever be the sample size, is not of the form $\sum_{i=1}^n T(X_i)$. In fact, the family of distributions in this example and the family $\mathcal{U}(0, \theta)$ are not exponential. Despite the existence of classes of examples such as these, starting with Koopman, Pitman, and Darmois, a theory has been built up that indicates that under suitable regularity conditions families of distributions, which admit k-dimensional sufficient statistics for all sample sizes, must be k-parameter exponential families. Some interesting results and a survey of the literature may be found in Brown (1986). Problem 1.6.10 is a special result of this type.

(1.6.24)

Summary. $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset R^k$, is a *k-parameter exponential family* of distributions if there are real-valued functions $\eta_1, \ldots, \eta_k$ and B on Θ, and real-valued functions $T_1, \ldots, T_k$, h on R^q such that the density (frequency) function of $P_{\boldsymbol{\theta}}$ can be written as

$$p(x, \boldsymbol{\theta}) = h(x) \exp[\sum_{j=1}^k \eta_j(\boldsymbol{\theta}) T_i(x) - B(\boldsymbol{\theta})], x \in \mathcal{X} \subset R^q. \tag{1.6.29}$$

(1.6.25)

$\mathbf{T}(X) = (T_1(X), \ldots, T_k(X))$ is called the *natural sufficient statistic* of the family. The *canonical k-parameter exponential family generated by $\mathbf{T}$ and h* is

(1.6.26)

$$q(x, \boldsymbol{\eta}) = h(x) \exp\{\mathbf{T}^T(x)\boldsymbol{\eta} - A(\boldsymbol{\eta})\}$$

where

(1.6.27)

$$A(\boldsymbol{\eta}) = \log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x) \exp\{\mathbf{T}^T(x)\boldsymbol{\eta}\} dx$$

in the continuous case, with integrals replaced by sums in the discrete case. The set

(1.6.28)

$$\mathcal{E} = \{\boldsymbol{\eta} \in R^k : -\infty < A(\boldsymbol{\eta}) < \infty\}$$

is called the *natural parameter space*. The set $\mathcal{E}$ is convex, the map $A : \mathcal{E} \to R$ is convex. If $\mathcal{E}$ has a nonempty interior in R^k and $\boldsymbol{\eta}_0 \in \mathcal{E}$, then $\mathbf{T}(X)$ has for $X \sim P_{\boldsymbol{\eta}_0}$ the moment-generating function

$$\psi(\mathbf{s}) = \exp\{A(\boldsymbol{\eta}_0 + \mathbf{s}) - A(\boldsymbol{\eta}_0)\}$$

for all $\mathbf{s}$ such that $\boldsymbol{\eta}_0 + \mathbf{s}$ is in $\mathcal{E}$. Moreover $E_{\boldsymbol{\eta}_0}[\mathbf{T}(X)] = \dot{A}(\boldsymbol{\eta}_0)$ and $Var_{\boldsymbol{\eta}_0}[T(X)] = \ddot{A}(\boldsymbol{\eta}_0)$ where $\dot{A}$ and $\ddot{A}$ denote the gradient and Hessian of A.

An exponential family is said to be of *rank* k if $\mathbf{T}$ is k-dimensional and $1, T_1, \ldots, T_k$ are linearly independent with positive P_θ probability for some $\theta \in \Theta$. If $\mathcal{P}$ is a canonical exponential family with $\mathcal{E}$ open, then the following are equivalent:

(i) $\mathcal{P}$ is of rank k,

(ii) η is identifiable,

(iii) $\mathrm{Var}_\eta(\mathbf{T})$ is positive definite,

(iv) the map $\eta \to \dot{A}(\eta)$ is 1 - 1 on $\mathcal{E}$,

(v) A is strictly convex on $\mathcal{E}$.

A family $\mathcal{F}$ of prior distributions for a parameter vector θ is called a *conjugate family* of priors to $p(x \mid \theta)$ if the posterior distribution of θ given $\mathbf{x}$ is a member of $\mathcal{F}$. The $(k+1)$-parameter exponential family

$$\pi_{\mathbf{t}}(\theta) = \exp\{\sum_{j=1}^{k} \eta_j(\theta) t_j - B(\theta) t_{k+1} - \log \omega\}$$

where

$$\omega = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\{\Sigma \eta_j(\theta) t_j - B(\theta)\} d\theta,$$

and

$$\mathbf{t} = (t_1, \ldots, t_{k+1}) \in \Omega = \{(t_1, \ldots, t_{k+1}) \in R^{k+1} : 0 < \omega < \infty\},$$

is conjugate to the exponential family $p(x|\theta)$ defined in (1.6.29).

1.7 PROBLEMS AND COMPLEMENTS

Problems for Section 1.1

1. Give a formal statement of the following models identifying the probability laws of the data and the parameter space. State whether the model in question is parametric or nonparametric.

(a) A geologist measures the diameters of a large number n of pebbles in an old stream bed. Theoretical considerations lead him to believe that the logarithm of pebble diameter is normally distributed with mean μ and variance σ^2. He wishes to use his observations to obtain some information about μ and σ^2 but has in advance no knowledge of the magnitudes of the two parameters.

(b) A measuring instrument is being used to obtain n independent determinations of a physical constant μ. Suppose that the measuring instrument is known to be biased to the positive side by 0.1 units. Assume that the errors are otherwise identically distributed normal random variables with known variance.