3 Exponential Families as a Unifier in Inference

Parametric inference is quite routinely used in relatively simple problems. Also, some problems are inherently parametric; for example, experiments that resemble iid sequences of a coin toss are automatically binomial experiments. The normal distribution is widely used in statistical practice; so is the Poisson. The Exponential family is a practically convenient unified family of distributions on finite dimensional Euclidean spaces that includes a large number of these standard parametric distributions as special cases. Specialized to the case of the real line, the Exponential family contains as special cases the normal, Poisson, Binomial, exponential, Gamma, negative binomial, etc.

However, there is much more to the Exponential family than just the fact that it includes many standard distributions as special cases. A number of important and useful calculations in statistical inference can be done in exact closed form all at one stroke within the framework of the Exponential family. As a matter of fact, if a parametric model is not in the Exponential family, we usually have to resort to asymptotic theory, because basic calculations would not be feasible in closed form for a given sample size n. Also, the Exponential family is the usual testing ground for the large spectrum of results in parametric statistical theory that require notions of *regularity*. Another attraction is that the unified calculations in the Exponential family setup have an element of mathematical neatness.

Distributions in the Exponential family have been used in classical statistics for decades. Recently, it has regained its historic importance in some novel inference problems that involve many parameters and a lot of data; the reason is that such problems are difficult to attack nonparametrically. A fundamental treatment of the general Exponential family is provided in this chapter. This unified treatment will save us repetitive and boring calculations for special distributions on a case by case basis. Classic expositions are available in Barndorff-Nielsen (1978), Brown (1986), and Lehmann and Casella (1998). Two other beautiful treatments are Bickel and Doksum (2006) and LeTac (1992). Liese and Miescke (2008) gives a rigorous modern treatment of Exponential families.

3.1 One Parameter Regular Exponential Family

Exponential families can have any finite number of parameters. For instance, as we will see, a normal distribution with a known mean is in the one parameter Exponential family, while a normal distribution with both parameters unknown is in the two parameter Exponential family. A bivariate normal distribution with all parameters unknown is in the five parameter Exponential family. As another example, if we take a normal distribution in which the mean and the variance are functionally related, e.g., the $N(\mu, \mu^2)$ distribution, then the distribution will be neither in the one parameter nor in the two parameter

Exponential family, but in a family called a *curved Exponential family.* We start with the one parameter regular Exponential family.

3.1.1 First Examples

Let us revisit an old example for simple illustration.

Example 3.1. (Normal Distribution with a Known Mean). Suppose $X \sim N(0, \sigma^2)$. Then the density of X is

$$f(x\,|\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} I_{x\in\mathcal{R}}.$$

This density is parametrized by a single parameter σ. Writing

$$\eta(\sigma) = -\frac{1}{2\sigma^2}, T(x) = x^2, \psi(\sigma) = \log\sigma, h(x) = \frac{1}{\sqrt{2\pi}} I_{x\in\mathcal{R}},$$

we can represent the density in the form

$$f(x\,|\sigma) = e^{\eta(\sigma)T(x)-\psi(\sigma)}h(x),$$

for any $\sigma \in \mathcal{R}_+$.

Next, suppose that we have an iid sample $X_1, X_2, \cdots, X_n \sim N(0, \sigma^2)$. Then the joint density of $X_1, X_2, \cdots, X_n$ is

$$f(x_1, x_2, \cdots, x_n\,|\sigma) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}} I_{x_1,x_2,\cdots,x_n\in\mathcal{R}}.$$

Now writing

$$\eta(\sigma) = -\frac{1}{2\sigma^2}, T(x_1, x_2, \cdots, x_n) = \sum_{i=1}^n x_i^2, \psi(\sigma) = n\log\sigma,$$

and

$$h(x_1, x_2, \cdots, x_n) = \frac{1}{(2\pi)^{n/2}} I_{x_1,x_2,\cdots,x_n\in\mathcal{R}},$$

once again we can represent the joint density in the same general form

$$f(x_1, x_2, \cdots, x_n\,|\sigma) = e^{\eta(\sigma)T(x_1,x_2,\cdots,x_n)-\psi(\sigma)}h(x_1, x_2, \cdots, x_n).$$

We notice that in this representation of the joint density $f(x_1, x_2, \cdots, x_n\,|\sigma)$, the statistic $T(X_1, X_2, \cdots, X_n)$ is still a one dimensional statistic, namely, $T(X_1, X_2, \cdots, X_n) = \sum_{i=1}^n X_i^2$. Using the fact that the sum of squares of n independent standard normal variables is a chi square variable with n degrees of freedom, we have that the density of $T(X_1, X_2, \cdots, X_n)$ is

$$f_T(t\,|\sigma) = \frac{e^{-\frac{t}{2\sigma^2}} t^{\frac{n}{2}-1}}{\sigma^n 2^{n/2}\Gamma(\frac{n}{2})} I_{t>0}.$$

This time, writing

$$\eta(\sigma) = -\frac{1}{2\sigma^2}, S(t) = t, \psi(\sigma) = n\log\sigma, h(t) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} I_{t>0},$$

once again we are able to write even the density of $T(X_1, X_2, \cdots, X_n) = \sum_{i=1}^n X_i^2$ in that same general form

$$f_T(t\,|\sigma) = e^{\eta(\sigma)S(t)-\psi(\sigma)}h(t).$$

Clearly, something very interesting is going on. We started with a basic density in a specific form, namely, $f(x\,|\sigma) = e^{\eta(\sigma)T(x)-\psi(\sigma)}h(x)$, and then we found that the joint density and the density of the relevant one dimensional statistic $\sum_{i=1}^n X_i^2$ in that joint density, are once again densities of exactly that same general form. It turns out that all of these phenomena are true of the entire family of densities which can be written in that general form, which is the one parameter Exponential family. Let us formally define it and we will then extend the definition to distributions with more than one parameter.

3.1.2 Definitions and Additional Examples

Definition 3.1. Let $X = (X_1, \cdots, X_d)$ be a d-dimensional random vector with a distribution $P_\theta, \theta \in \Theta \subseteq \mathcal{R}$.

Suppose $X_1, \cdots, X_d$ are jointly continuous. The family of distributions $\{P_\theta, \theta \in \Theta\}$ is said to belong to the *one parameter Exponential family* if the density of $X = (X_1, \cdots, X_d)$ may be represented in the form

$$f(x\,|\theta) = e^{\eta(\theta)T(x)-\psi(\theta)}h(x),$$

for some real valued functions $T(x), \psi(\theta)$ and $h(x) \geq 0$.

If $X_1, \cdots, X_d$ are jointly discrete, then $\{P_\theta, \theta \in \Theta\}$ is said to belong to the one parameter Exponential family if the joint pmf $p(x\,|\theta) = P_\theta(X_1 = x_1, \cdots, X_d = x_d)$ may be written in the form

$$p(x\,|\theta) = e^{\eta(\theta)T(x)-\psi(\theta)}h(x),$$

for some real valued functions $T(x), \psi(\theta)$ and $h(x) \geq 0$.

Note that the functions η, T and h are not unique. For example, in the product ηT, we can multiply T by some constant c and divide η by it. Similarly, we can play with constants in the function h.

Definition 3.2. Suppose $X = (X_1, \cdots, X_d)$ has a distribution $P_\theta, \theta \in \Theta$, belonging to the one parameter Exponential family. Then the statistic $T(X)$ is called *the natural sufficient statistic* for the family $\{P_\theta\}$.

The notion of a sufficient statistic is a fundamental one in statistical theory and its applications. A sufficient statistic is supposed to contain by itself all of the information about the unknown parameters of the underlying distribution that the entire sample could have provided. Reduction by sufficiency in widely used models usually makes just simple common sense. We will come back to the issue of sufficiency once again in Chapter 7.

We will now see examples of a few more common distributions that belong to the one parameter Exponential family.

Example 3.2. (Binomial Distribution). Let $X \sim Bin(n,p)$, with $n \geq 1$ considered as known, and $0 < p < 1$ a parameter. We represent the pmf of X in the one parameter Exponential family form.

$$f(x\,|p) = \binom{n}{x} p^x (1-p)^{n-x} I_{\{x\in\{0,1,\cdots,n\}\}} = \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1-p)^n I_{\{x\in\{0,1,\cdots,n\}\}}$$

$$= \binom{n}{x} e^{x \log \frac{p}{1-p} + n\log(1-p)} I_{\{x\in\{0,1,\cdots,n\}\}}.$$

Writing $\eta(p) = \log \frac{p}{1-p}, T(x) = x, \psi(p) = -n\log(1-p)$, and $h(x) = \binom{n}{x} I_{\{x\in\{0,1,\cdots,n\}\}}$, we have represented the pmf $f(x\,|p)$ in the one parameter Exponential family form, as long as $p \in (0,1)$. For $p = 0$ or 1, the distribution becomes a one point distribution. Consequently, the family of distributions $\{f(x\,|p), 0 < p < 1\}$ forms a one parameter Exponential family, but if either of the boundary values $p = 0, 1$ is included, the family is not in the Exponential family.

Example 3.3. (Normal Distribution with a Known Variance). Suppose $X \sim N(\mu, \sigma^2)$, where σ is considered known, and $\mu \in \mathcal{R}$ a parameter. Then,

$$f(x\,|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2} + \mu x - \frac{\mu^2}{2}} I_{x\in\mathcal{R}},$$

which can be written in the one parameter Exponential family form by witing $\eta(\mu) = \mu, T(x) = x, \psi(\mu) = \frac{\mu^2}{2}$, and $h(x) = e^{-\frac{x^2}{2}} I_{x\in\mathcal{R}}$. So, the family of distributions $\{f(x\,|\mu), \mu \in \mathcal{R}\}$ forms a one parameter Exponential family.

Example 3.4. (Gamma Distribution). Suppose X has the Gamma density $\frac{e^{-\frac{x}{\lambda}} x^{\alpha-1}}{\lambda^\alpha \Gamma(\alpha)} I_{x>0}$. As such, it has two parameters λ, α. If we assume that α is known, then we may write the density in the one parameter Exponential family form:

$$f(x\,|\lambda) = e^{-\frac{x}{\lambda} - \alpha\log\lambda} \frac{x^{\alpha-1}}{\Gamma(\alpha)} I_{x>0},$$

and recognize it as a density in the Exponential family with $\eta(\lambda) = -\frac{1}{\lambda}, T(x) = x, \psi(\lambda) = \alpha\log\lambda, h(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)} I_{x>0}$.

If we assume that λ is known, once again, by writing the density as

$$f(x\,|\alpha) = e^{\alpha\log x - \alpha(\log\lambda) - \log\Gamma(\alpha)} e^{-\frac{x}{\lambda}} I_{x>0},$$

we recognize it as a density in the Exponential family with $\eta(\alpha) = \alpha, T(x) = \log x, \psi(\alpha) = \alpha(\log\lambda) + \log\Gamma(\alpha), h(x) = e^{-\frac{x}{\lambda}} I_{x>0}$.

Example 3.5. (An Unusual Gamma Distribution). Suppose we have a Gamma density in which the mean is known, say, $E(X) = 1$. This means that $\alpha\lambda = 1 \Rightarrow \lambda = \frac{1}{\alpha}$. Parametrizing the density with α, we have

$$f(x\,|\alpha) = e^{-\alpha x + \alpha \log x} \frac{\alpha^\alpha}{\Gamma(\alpha)} \frac{1}{x} I_{x>0}$$

$$= e^{\alpha \left[\log x - x\right] - \left[\log \Gamma(\alpha) - \alpha \log \alpha\right]} \frac{1}{x} I_{x>0},$$

which is once again in the one parameter Exponential family form with $\eta(\alpha) = \alpha, T(x) = \log x - x, \psi(\alpha) = \log \Gamma(\alpha) - \alpha \log \alpha, h(x) = \frac{1}{x} I_{x>0}$.

Example 3.6. (A Normal Distribution Truncated to a Set). Suppose a certain random variable W has a normal distribution with mean μ and variance one. We saw in Example 18.3 that this is in the one parameter Exponential family. Suppose now that the variable W can be physically observed only when its value is inside some set A. For instance, if $W > 2$, then our measuring instruments cannot tell what the value of W is. In such a case, the variable X that is truly observed has a normal distribution truncated to the set A. For simplicity, take A to be $A = [a, b]$, an interval. Then, the density of X is

$$f(x\,|\mu) = \frac{e^{-\frac{(x-\mu)^2}{2}}}{\sqrt{2\pi}[\Phi(b-\mu) - \Phi(a-\mu)]} I_{a \le x \le b}.$$

This can be written as

$$f(x\,|\mu) = \frac{1}{\sqrt{2\pi}} e^{\mu x - \frac{\mu^2}{2} - \log\left[\Phi(b-\mu) - \Phi(a-\mu)\right]} e^{-\frac{x^2}{2}} I_{a \le x \le b},$$

and we recognize this to be in the Exponential family form with $\eta(\mu) = \mu, T(x) = x, \psi(\mu) = \frac{\mu^2}{2} + \log[\Phi(b-\mu) - \Phi(a-\mu)]$, and $h(x) = e^{-\frac{x^2}{2}} I_{a \le x \le b}$. Thus, the distribution of W truncated to $A = [a, b]$ is still in the one parameter Exponential family. This phenomenon is in fact more general. Ex **(Some Distributions not in the Exponential Family).** It is clear from the definition of a one parameter Exponential family that if a certain family of distributions $\{P_\theta, \theta \in \Theta\}$ belongs to the one parameter Exponential family, then each P_θ has exactly the same support. Precisely, for any fixed $\theta, P_\theta(A) > 0$ if and only if $\int_A h(x)dx > 0$, and in the discrete case, $P_\theta(A) > 0$ if and only if $A \cap \mathcal{X} \neq \emptyset$, where $\mathcal{X}$ is the countable set $\mathcal{X} = \{x : h(x) > 0\}$. As a consequence of this common support fact, the so called *irregular distributions* whose support depends on the parameter cannot be members of the Exponential family. Examples would be the family of $U[0, \theta], U[-\theta, \theta]$ distributions, etc. Likewise, the *shifted Exponential density* $f(x\,|\theta) = e^{\theta - x} I_{x>\theta}$ cannot be in the Exponential family.

Some other common distributions are also not in the Exponential family, but for other reasons. An important example is the family of Cauchy distributions given by the location parameter form $f(x\,|\mu) = \frac{1}{\pi[1+(x-\mu)^2]} I_{x\in\mathcal{R}}$. Suppose that it is. Then, we can find functions $\eta(\mu), T(x)$ such that for all x, μ,

$$e^{\eta(\mu)T(x)} = \frac{1}{1+(x-\mu)^2} \Rightarrow \eta(\mu)T(x) = -\log(1+(x-\mu)^2)$$

$$\Rightarrow \eta(0)T(x) = -\log(1+x^2) \Rightarrow T(x) = -c\log(1+x^2)$$

for some constant c.
Plugging this back, we get, for all x, μ,

$$-c\eta(\mu)\log(1+x^2) = -\log(1+(x-\mu)^2) \Rightarrow \eta(\mu) = \frac{1}{c}\frac{\log(1+(x-\mu)^2)}{\log(1+x^2)}.$$

This means that $\frac{\log(1+(x-\mu)^2)}{\log(1+x^2)}$ must be a constant function of x, which is a contradiction. The choice of $\mu = 0$ as the special value of μ is not important.

3.1.3 Canonical Form and General Properties

Suppose $\{P_\theta, \theta \in \Theta\}$ is a family belonging to the one parameter Exponential family, with density (or pmf) of the form $f(x\,|\theta) = e^{\eta(\theta)T(x)-\psi(\theta}h(x)$. If $\eta(\theta)$ is a one-to-one function of θ, then we can drop θ altogether, and parametrize the distribution in terms of η itself. If we do that, we get a reparametrized density g in the form $e^{\eta T(x)-\psi^*(\eta)}h(x)$. By a slight abuse of notation, we will again use the notation f for g and ψ for ψ^*.

Definition 3.3. Let $X = (X_1, \cdots, X_d)$ have a distribution $P_\eta, \eta \in \mathcal{T} \subseteq \mathcal{R}$. The family of distributions $\{P_\eta, \eta \in \mathcal{T}\}$ is said to belong to the *canonical one parameter Exponential family* if the density (pmf) of P_η may be written in the form

$$f(x\,|\eta) = e^{\eta T(x)-\psi(\eta)}h(x),$$

where

$$\eta \in \mathcal{T} = \{\eta : e^{\psi(\eta)} = \int_{\mathcal{R}^d} e^{\eta T(x)}h(x)dx < \infty\},$$

in the continuous case, and

$$\mathcal{T} = \{\eta : e^{\psi(\eta)} = \sum_{x\in\mathcal{X}} e^{\eta T(x)}h(x) < \infty\},$$

in the discrete case, with $\mathcal{X}$ being the countable set on which $h(x) > 0$.
For a distribution in the canonical one parameter Exponential family, the parameter η is called the *natural parameter*, and $\mathcal{T}$ is called *the natural parameter space*. Note that $\mathcal{T}$ describes the largest set of values of η for which the density (pmf) can be defined. In a particular application, we may have extraneous knowledge that η belongs to some proper

subset of $\mathcal{T}$. hus, $\{P_\eta\}$ with $\eta \in \mathcal{T}$ is called the *full canonical one parameter Exponential family*. We generally refer to the full family, unless otherwise stated.
The canonical Exponential family is called *regular* if $\mathcal{T}$ is an open set in $\mathcal{R}$, and it is called *nonsingular* if $\text{Var}_\eta(T(X)) > 0$ for all $\eta \in \mathcal{T}^0$, the interior of the natural parameter space $\mathcal{T}$.

It is analytically convenient to work with an Exponential family distribution in its canonical form. Once a result has been derived for the canonical form, if desired we can rewrite the answer in terms of the original parameter θ. Doing this retransformation at the end is algebraically and notationally simpler than carrying the original function $\eta(\theta)$ and often its higher derivatives with us throughout a calculation. Most of our formulae and theorems below will be given for the canonical form.

Example 3.7. (Binomial Distribution in Canonical Form). Let $X \sim Bin(n,p)$ with the pmf $\binom{n}{x}p^x(1-p)^{n-x}I_{x\in\{0,1,\cdots,n\}}$. In Example 18.2, we represented this pmf in the Exponential family form

$$f(x\,|p) = e^{x\log\frac{p}{1-p} - nlog(1-p)}\binom{n}{x}I_{x\in\{0,1,\cdots,n\}}.$$

If we write $\log\frac{p}{1-p} = \eta$, then $\frac{p}{1-p} = e^\eta$, and hence, $p = \frac{e^\eta}{1+e^\eta}$, and $1-p = \frac{1}{1+e^\eta}$. Therefore, the canonical Exponential family form of the binomial distribution is

$$f(x\,|\eta) = e^{\eta x - n\log(1+e^\eta)}\binom{n}{x}I_{x\in\{0,1,\cdots,n\}},$$

and the natural parameter space is $\mathcal{T} = \mathcal{R}$.

3.2 Multiparameter Exponential Family

Similar to the case of distributions with only one parameter, several common distributions with multiple parameters also belong to a general multiparameter Exponential family. An example is the normal distribution on $\mathcal{R}$ with both parameters unknown. Another example is a multivariate normal distribution. Analytic techniques and properties of multiparameter Exponential families are very similar to those of the one parameter Exponential family. Because of that reason, most of our presentation in this section dwells on examples.

Definition 3.4. Let $X = (X_1, \cdots, X_d)$ have a distribution $P_\theta, \theta \in \Theta \subseteq \mathcal{R}^k$. The family of distributions $\{P_\theta, \theta \in \Theta\}$ is said to belong to the k-parameter Exponential family if its density (pmf) may be represented in the form

$$f(x\,|\theta) = e^{\sum_{i=1}^k \eta_i(\theta)T_i(x) - \psi(\theta)}h(x).$$

Again, obviously, the choice of the relevant functions η_i, T_i, h is not unique. As in the one parameter case, the vector of statistics $(T_1, \cdots, T_k)$ is called the natural sufficient statistic, and if we reparametrize by using $\eta_i = \eta_i(\theta), i = 1, 2, \cdots, k$, the family is called the k-parameter canonical Exponential family.

There is an implicit assumption in this definition that the number of *freely varying* θ's is the same as the number of freely varying η's, and that these are both equal to the specific k in the context. The formal way to say this is to assume the following:

Assumption The dimension of Θ as well as the dimension of the image of Θ under the map $(\theta_1, \theta_2, \cdots, \theta_k) \longrightarrow (\eta_1(\theta_1, \theta_2, \cdots, \theta_k), \eta_2(\theta_1, \theta_2, \cdots, \theta_k), \cdots, \eta_k(\theta_1, \theta_2, \cdots, \theta_k))$ are equal to k.

There are some important examples where this assumption does not hold. They will not be counted as members of a k-parameter Exponential family. The name curved Exponential family is commonly used for them, and this will be discussed in the supplementary section of this chapter.

The terms *canonical form, natural parameter, and natural parameter space* will mean the same things as in the one parameter case. Thus, if we parametrize the distributions by using $\eta_1, \eta_2, \cdots, \eta_k$ as the k parameters, then the vector $\eta = (\eta_1, \eta_2, \cdots, \eta_k)$ is called the natural parameter vector, the parametrization $f(x\,|\eta) = e^{\sum_{i=1}^k \eta_i T_i(x) - \psi(\eta)} h(x)$ is called the canonical form, and the set of all vectors η for which $f(x\,|\eta)$ is a valid density (pmf) is called the natural parameter space. The main theorems for the case $k = 1$ hold for a general k.

Theorem 3.1. The results of Theorem 5.1 and 5.5 hold for the k-parameter Exponential family.

The proofs are almost verbatim the same. The moment formulas differ somewhat due to the presence of more than one parameter in the current context.

Theorem 3.2. Suppose $X = (X_1, \cdots, X_d)$ has a distribution $P\eta, \eta \in \mathcal{T}$, belonging to the canonical k-parameter Exponential family, with a density (pmf)

$$f(x\,|\eta) = e^{\sum_{i=1}^k \eta_i T_i(x) - \psi(\eta)} h(x),$$

where

$$\mathcal{T} = \{\eta \in \mathcal{R}^k : \int_{\mathcal{R}^d} e^{\sum_{i=1}^k \eta_i T_i(x)} h(x) dx < \infty\}$$

(and the integral being replaced by a sum in the discrete case).

(a) At any $\eta \in \mathcal{T}^0$,

$$e^{\psi(\eta)} = \int_{\mathcal{R}^d} e^{\sum_{i=1}^k \eta_i T_i(x)} h(x) dx$$

is infinitely partially differentiable with respect to each η_i, nd the partial derivatives of any order can be obtained by differentiating inside the integral sign.

$$(b)\; E_\eta[T_i(X)] = \frac{\partial}{\partial \eta_i}\psi(\eta);\; \mathrm{Cov}_\eta(T_i(X), T_j(X)) = \frac{\partial^2}{\partial \eta_i \partial \eta_j}\psi(\eta), 1 \le i, j \le k.$$

(c) If η, t are such that $\eta, \eta + t \in \mathcal{T}$, then the joint mgf of $(T_1(X), \cdots, T_k(X))$ exists and equals

$$M_\eta(t) = e^{\psi(\eta+t)-\psi(\eta)}.$$

An important new terminology is that of a *full rank*.

Definition 3.5. A family of distributions $\{P\eta, \eta \in \mathcal{T}\}$ belonging to the canonical k-parameter Exponential family is called full rank if at every $\eta \in \mathcal{T}^0$, the $k \times k$ covariance matrix $\left(\left(\frac{\partial^2}{\partial\eta_i\partial\eta_j}\psi(\eta)\right)\right)$ is nonsingular.

Definition 3.6. (Fisher Information Matrix). Suppose a family of distributions in the canonical k-parameter Exponential family is nonsingular. Then, for $\eta \in \mathcal{T}^0$, the matrix $\left(\left(\frac{\partial^2}{\partial\eta_i\partial\eta_j}\psi(\eta)\right)\right)$ is called the Fisher information matrix (at η).

The Fisher information matrix is of paramount importance in parametric statistical theory and lies at the heart of finite and large sample optimality theory in statistical inference problems for general regular parametric families.

We will now see some examples of distributions in k-parameter Exponential families where $k > 1$.

Example 3.8. (Two Parameter Normal Distribution). Suppose $X \sim N(\mu, \sigma^2)$, and we consider both μ, σ to be parameters. If we denote $(\mu, \sigma) = (\theta_1, \theta_2) = \theta$, then parametrized by θ, the density of X is

$$f(x\,|\theta) = \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{(x-\theta_1)^2}{2\theta_2^2}} I_{x\in\mathcal{R}} = \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{x^2}{2\theta_2^2}+\frac{\theta_1 x}{\theta_2^2}-\frac{\theta_1^2}{2\theta_2^2}} I_{x\in\mathcal{R}}.$$

This is in the two parameter Exponential family with

$$\eta_1(\theta) = -\frac{1}{2\theta_2^2}, \eta_2(\theta) = \frac{\theta_1}{\theta_2^2}, T_1(x) = x^2, T_2(x) = x,$$

$$\psi(\theta) = \frac{\theta_1^2}{2\theta_2^2} + \log\theta_2, h(x) = \frac{1}{\sqrt{2\pi}} I_{x\in\mathcal{R}}.$$

The parameter space in the θ parametrization is

$$\Theta = (-\infty, \infty) \otimes (0, \infty).$$

If we want the canonical form, we let $\eta_1 = -\frac{1}{2\theta_2^2}, \eta_2 = \frac{\theta_1}{\theta_2^2}$, and $\psi(\eta) = -\frac{\eta_2^2}{4\eta_1} - \frac{1}{2}\log(-\eta_1)$. The natural parameter space for (η_1, η_2) is $(-\infty, 0) \otimes (-\infty, \infty)$.

Example 3.9. (Two Parameter Gamma). It was seen in Example 5.4 that if we fix one of the two parameters of a Gamma distribution, then it becomes a member of the one parameter Exponential family. We show in this example that the general Gamma distribution is a member of the two parameter Exponential family. To show this, just observe that with $\theta = (\alpha, \lambda) = (\theta_1, \theta_2)$,

$$f(x\,|\theta) = e^{-\frac{x}{\theta_2}+\theta_1 \log x - \theta_1 \log \theta_2 - \log \Gamma(\theta_1)} \frac{1}{x} I_{x>0}.$$

This is in the two parameter Exponential family with $\eta_1(\theta) = -\frac{1}{\theta_2}, \eta_2(\theta) = \theta_1, T_1(x) = x, T_2(x) = \log x, \psi(\theta) = \theta_1 \log \theta_2 + \log \Gamma(\theta_1)$, and $h(x) = \frac{1}{x} I_{x>0}$. The parameter space in the θ-parametrization is $(0, \infty) \otimes (0, \infty)$. For the canonical form, use $\eta_1 = -\frac{1}{\theta_2}, \eta_2 = \theta_1$, and so, the natural parameter space is $(-\infty, 0) \otimes (0, \infty)$. The natural sufficient statistic is $(X, \log X)$.

Example 3.10. (The General Multivariate Normal Distribution). Suppose $X \sim N_d(\mu, \Sigma)$, where μ is arbitrary and Σ is positive definite (and of course, symmetric). Writing $\theta = (\mu, \Sigma)$, we can think of θ as a subset in an Euclidean space of dimension

$$k = d + d + \frac{d^2 - d}{2} = d + \frac{d(d+1)}{2} = \frac{d(d+3)}{2}.$$

The density of X is

$$f(x\,|\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)} I_{x \in \mathcal{R}^d}.$$

$$= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}x'\Sigma^{-1}x + \mu'\Sigma^{-1}x - \frac{1}{2}\mu'\Sigma^{-1}\mu} I_{x \in \mathcal{R}^d}$$

$$= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}\sum\sum_{i,j} \sigma^{ij} x_i x_j + \sum_i (\sum_k \sigma^{ki}\mu_k) x_i - \frac{1}{2}\mu'\Sigma^{-1}\mu} I_{x \in \mathcal{R}^d}$$

$$= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}\sum_i \sigma^{ii} x_i^2 - \sum\sum_{i<j} \sigma^{ij} x_i x_j + \sum_i (\sum_k \sigma^{ki}\mu_k) x_i - \frac{1}{2}\mu'\Sigma^{-1}\mu} I_{x \in \mathcal{R}^d}.$$

We have thus represented the density of X in the k-parameter Exponential family form with the k-dimensional natural sufficient statistic

$$T(X) = (X_1, \cdots, X_d, X_1^2, \cdots, X_d^2, X_1 X_2, \cdots, X_{d-1} X_d),$$

and the natural parameters defined by

$$\sum_k \sigma^{k1}\mu_k, \cdots, \sum_k \sigma^{kd}\mu_k, -\frac{1}{2}\sigma^{11}, \cdots, -\frac{1}{2}\sigma^{dd}, -\sigma^{12}, \cdots, -\sigma^{d-1,d}.$$

Example 3.11. (Multinomial Distribution). Consider the $k+1$ cell multinomial distribution with cell probabilities $p_1, p_2, \cdots, p_k, p_{k+1} = 1 - \sum_{i=1}^k p_i$. Writing $\theta = (p_1, p_2, \cdots, p_k)$, the joint pmf of $X = (X_1, X_2, \cdots, X_k)$, the cell frequencies of the first k cells, is

$$f(x\,|\theta) = \frac{n!}{(\prod_{i=1}^k x_i!)(n - \sum_{i=1}^k x_i)!} \prod_{i=1}^k p_i^{x_i} (1 - \sum_{i=1}^k p_i)^{n - \sum_{i=1}^k x_i} I_{x_1, \cdots, x_k \geq 0, \sum_{i=1}^k x_i \leq n}$$

$$= \frac{n!}{(\prod_{i=1}^k x_i!)(n - \sum_{i=1}^k x_i)!} e^{\sum_{i=1}^k (\log p_i) x_i - \log(1-\sum_{i=1}^k p_i)(\sum_{i=1}^k x_i) + n \log(1-\sum_{i=1}^k p_i)} I_{x_1,\cdots,x_k \ge 0, \sum_{i=1}^k x_i \le n}$$

$$= \frac{n!}{(\prod_{i=1}^k x_i!)(n - \sum_{i=1}^k x_i)!} e^{\sum_{i=1}^k (\log \frac{p_i}{1-\sum_{i=1}^k p_i}) x_i + n \log(1-\sum_{i=1}^k p_i)} I_{x_1,\cdots,x_k \ge 0, \sum_{i=1}^k x_i \le n}.$$

This is in the k-parameter Exponential family form with the natural sufficient statistic and natural parameters

$$T(X) = (X_1, X_2, \cdots, X_k),\ \eta_i = \log \frac{p_i}{1 - \sum_{i=1}^k p_i}, 1 \le i \le k.$$

Example 3.12. (Two Parameter Inverse Gaussian Distribution). The general inverse Gaussian density (see Chapter 2) is given by

$$f(x\,|\theta_1, \theta_2) = \left(\frac{\theta_2}{\pi x^3}\right)^{1/2} e^{-\theta_1 x - \frac{\theta_2}{x} + 2\sqrt{\theta_1 \theta_2}} I_{x>0};$$

the parameter space for $\theta = (\theta_1, \theta_2)$ is $[0, \infty) \otimes (0, \infty)$. The general inverse Gaussian density $f(x\,|\theta_1, \theta_2)$ is used as a model for skewed densities, and interestingly, also arises in some problems of *random walks* in probability in a fundamental way.

It is clear from the formula for $f(x\,|\theta_1, \theta_2)$ that it is a member of the two parameter Exponential family with the natural sufficient statistic $T(X) = (X, \frac{1}{X})$ and the natural parameter space $\mathcal{T} = (-\infty, 0] \otimes (-\infty, 0)$. Note that the natural parameter space is not open.

3.3 At Instructor's Discretion

3.3.1 Convexity Properties

Written in its canonical form, a density (pmf) in an Exponential family has some convexity properties. These convexity properties are useful in manipulating with moments and other functionals of $T(X)$, the natural sufficient statistic appearing in the expression for the density of the distribution.

Theorem 3.3. The natural parameter space $\mathcal{T}$ is convex, and $\psi(\eta)$ is a convex function on $\mathcal{T}$.

Proof: We consider the continuous case only, as the discrete case admits basically the same proof. Let η_1, η_2 be two members of $\mathcal{T}$, and let $0 < \alpha < 1$. We need to show that $\alpha\eta_1 + (1-\alpha)\eta_2$ belongs to $\mathcal{T}$, i.e.,

$$\int_{\mathcal{R}^d} e^{(\alpha\eta_1 + (1-\alpha)\eta_2)T(x)} h(x) dx < \infty.$$

But,

$$\int_{\mathcal{R}^d} e^{(\alpha\eta_1 + (1-\alpha)\eta_2)T(x)} h(x) dx = \int_{\mathcal{R}^d} e^{\alpha\eta_1 T(x)} \times e^{(1-\alpha)\eta_2 T(x)} h(x) dx$$

$$= \int_{\mathcal{R}^d} \left(e^{\eta_1 T(x)}\right)^{\alpha} \left(e^{\eta_2 T(x)}\right)^{1-\alpha} h(x)dx$$

$$\leq \left(\int_{\mathcal{R}^d} e^{\eta_1 T(x)} h(x)dx\right)^{\alpha} \left(\int_{\mathcal{R}^d} e^{\eta_2 T(x)} h(x)dx\right)^{1-\alpha}$$

(by Holder's inequality)

$$< \infty,$$

because, by hypothesis, $\eta_1, \eta_2 \in \mathcal{T}$, and hence, $\int_{\mathcal{R}^d} e^{\eta_1 T(x)} h(x)dx$, and $\int_{\mathcal{R}^d} e^{\eta_2 T(x)} h(x)dx$ are both finite.

Note that in this argument, we have actually proved the inequality

$$e^{\psi(\alpha\eta_1 + (1-\alpha)\eta_2)} \leq e^{\alpha\psi(\eta_1) + (1-\alpha)\psi(\eta_2)}.$$

But this is the same as saying

$$\psi(\alpha\eta_1 + (1-\alpha)\eta_2) \leq \alpha\psi(\eta_1) + (1-\alpha)\psi(\eta_2),$$

i.e., $\psi(\eta)$ is a convex function on $\mathcal{T}$. ♣

3.3.2 Moments and Moment Generating Function

The next result is a very special fact about the canonical Exponential family, and is the source of a large number of closed form formulas valid for the entire canonical Exponential family. The fact itself is actually a fact in mathematical analysis. Due to the special form of Exponential family densities, the fact in analysis translates to results for the Exponential family, an instance of interplay between mathematics and statistics and probability.

Theorem 3.4. (a) The function $e^{\psi(\eta)}$ is infinitely differentiable at every $\eta \in \mathcal{T}^0$. Furthermore, in the continuous case, $e^{\psi(\eta)} = \int_{\mathcal{R}^d} e^{\eta T(x)} h(x)dx$ can be differentiated any number of times inside the integral sign, and in the discrete case, $e^{\psi(\eta)} = \sum_{x \in \mathcal{X}} e^{\eta T(x)} h(x)$ can be differentiated any number of times inside the sum.

(b) In the continuous case, for any $k \geq 1$,

$$\frac{d^k}{d\eta^k} e^{\psi(\eta)} = \int_{\mathcal{R}^d} [T(x)]^k e^{\eta T(x)} h(x)dx,$$

and in the discrete case,

$$\frac{d^k}{d\eta^k} e^{\psi(\eta)} = \sum_{x \in \mathcal{X}} [T(x)]^k e^{\eta T(x)} h(x).$$

Proof: Take $k = 1$. Then, by the definition of derivative of a function, $\frac{d}{d\eta} e^{\psi(\eta)}$ exists if and only if $\lim_{\delta \to 0} [\frac{e^{\psi(\eta+\delta)} - e^{\psi(\eta)}}{\delta}]$ exists. But,

$$\frac{e^{\psi(\eta+\delta)} - e^{\psi(\eta)}}{\delta} = \int_{\mathcal{R}^d} \frac{e^{(\eta+\delta)T(x)} - e^{\eta T(x)}}{\delta} h(x)dx,$$

and by an application of the *Dominated convergence theorem*, $\lim_{\delta\to 0}\int_{\mathcal{R}^d}\frac{e^{(\eta+\delta)T(x)}-e^{\eta T(x)}}{\delta}h(x)dx$ exists, and the limit can be carried inside the integral, to give

$$\lim_{\delta\to 0}\int_{\mathcal{R}^d}\frac{e^{(\eta+\delta)T(x)}-e^{\eta T(x)}}{\delta}h(x)dx=\int_{\mathcal{R}^d}\lim_{\delta\to 0}\frac{e^{(\eta+\delta)T(x)}-e^{\eta T(x)}}{\delta}h(x)dx$$
$$=\int_{\mathcal{R}^d}\frac{d}{d\eta}e^{\eta T(x)}h(x)dx=\int_{\mathcal{R}^d}T(x)e^{\eta T(x)}h(x)dx.$$

Now use induction on k by using the Dominated convergence theorem again. ♣

This compact formula for an arbitrary derivative of $e^{\psi(\eta)}$ leads to the following important moment formulas.

Theorem 3.5. At any $\eta\in\mathcal{T}^0$,

$$(a)\; E_\eta[T(X)]=\psi'(\eta);\;\; \mathrm{Var}_\eta[T(X)]=\psi''(\eta);$$

(b) The coefficients of skewness and kurtosis of $T(X)$ equal

$$\beta_(\eta)=\frac{\psi^{(3)}(\eta)}{[\psi''(\eta)]^{3/2}};\text{ and }\gamma(\eta)=\frac{\psi^{(4)}(\eta)}{[\psi''(\eta)]^2};$$

(c) At any t such that $\eta+t\in\mathcal{T}$, the mgf of $T(X)$ exists and equals

$$M_\eta(t)=e^{\psi(\eta+t)-\psi(\eta)}.$$

Proof: Again, we take just the continuous case. Consider the result of the previous theorem that for any $k\geq 1$, $\frac{d^k}{d\eta^k}e^{\psi(\eta)}=\int_{\mathcal{R}^d}[T(x)]^k e^{\eta T(x)}h(x)dx$. Using this for $k=1$, we get

$$\psi'(\eta)e^{\psi(\eta)}=\int_{\mathcal{R}^d}T(x)e^{\eta T(x)}h(x)dx\Rightarrow\int_{\mathcal{R}^d}T(x)e^{\eta T(x)-\psi(\eta)}h(x)dx=\psi'(\eta),$$

which gives the result $E_\eta[T(X)]=\psi'(\eta)$.

Similarly,

$$\frac{d^2}{d\eta^2}e^{\psi(\eta)}=\int_{\mathcal{R}^d}[T(x)]^2e^{\eta T(x)}h(x)dx\Rightarrow[\psi''(\eta)+\{\psi'(\eta)\}^2]e^{\psi(\eta)}=\int_{\mathcal{R}^d}[T(x)]^2e^{\eta T(x)}h(x)dx$$
$$\Rightarrow\psi''(\eta)+\{\psi'(\eta)\}^2=\int_{\mathcal{R}^d}[T(x)]^2e^{\eta T(x)-\psi(\eta)}h(x)dx,$$

which gives $E_\eta[T(X)]^2=\psi''(\eta)+\{\psi'(\eta)\}^2$. Combine this with the already obtained result that $E_\eta[T(X)]=\psi'(\eta)$, and we get $\mathrm{Var}_\eta[T(X)]=E_\eta[T(X)]^2-(E_\eta[T(X)])^2=\psi''(\eta)$.

The coefficient of skewness is defined as

$$\beta_\eta=\frac{E[T(X)-ET(X)]^3}{(\mathrm{Var}T(X))^{3/2}}.$$

To obtain

$$E[T(X)-ET(X)]^3=E[T(X)]^3-3E[T(X)]^2E[T(X)]+2[ET(X)]^3,$$

use the identity

$$\frac{d^3}{d\eta^3}e^{\psi(\eta)} = \int_{\mathcal{R}^d}[T(x)]^3 e^{\eta T(x)}h(x)dx.$$

Then use the fact that the third derivative of $e^{\psi(\eta)}$ is $e^{\psi(\eta)}\Big[\psi^{(3)}(\eta) + 3\psi'(\eta)\psi''(\eta) + \{\psi'(\eta)\}^3\Big]$. As we did in our proofs for the mean and the variance above, transfer $e^{\psi(\eta)}$ into the integral on the right hand side and then simplify. This will give $E[T(X) - ET(X)]^3 = \psi^{(3)}(\eta)$, and the skewness formula follows. The formula for kurtosis is proved by the same argument, using $k = 4$ in the derivative identity

$$\frac{d^k}{d\eta^k}e^{\psi(\eta)} = \int_{\mathcal{R}^d}[T(x)]^k e^{\eta T(x)}h(x)dx$$

Finally, for the mgf formula,

$$M_\eta(t) = E_\eta[e^{tT(X)}] = \int_{\mathcal{R}^d} e^{tT(X)}e^{\eta T(x)-\psi(\eta)}h(x)dx = e^{-\psi(\eta)}\int_{\mathcal{R}^d} e^{(t+\eta)T(x)}h(x)dx$$

$$= e^{-\psi(\eta)}e^{\psi(t+\eta)}\int_{\mathcal{R}^d} e^{(t+\eta)T(x)-\psi(t+\eta)}h(x)dx = e^{-\psi(\eta)}e^{\psi(t+\eta)} \times 1$$

$$= e^{\psi(t+\eta)-\psi(\eta)}.$$

An important consequence of the mean and the variance formulas is the following monotonicity result. ♣

Corollary 3.1. For a nonsingular canonical Exponential family, $E_\eta[T(X)]$ is strictly increasing in η on $\mathcal{T}^0$.

Proof: From part (a) of Theorem 18.3, the variance of $T(X)$ is the derivative of the expectation of $T(X)$, and by nonsingularity, the variance is strictly positive. This implies that the expectation is strictly increasing.

As a consequence of this strict monotonicity of the mean of $T(X)$ in the natural parameter, nonsingular canonical Exponential families may be reparametrized by using the mean of T itself as the parameter. This is useful for some purposes.

Example 3.13. (Binomial Distribution). From Example 18.9, in the canonical representation of the binomial distribution, $\psi(\eta) = n\log(1+e^\eta)$. By direct differentiation,

$$\psi'(\eta) = \frac{ne^\eta}{1+e^\eta};\ \psi''(\eta) = \frac{ne^\eta}{(1+e^\eta)^2};$$

$$\psi^{(3)}(\eta) = \frac{-ne^\eta(e^\eta - 1)}{(1+e^\eta)^3};\ \psi^{(4)}(\eta) = \frac{ne^\eta(e^{2\eta} - 4e^\eta + 1)}{(1+e^\eta)^4}.$$

Now recall from Example 5.7 that the success probability p and the natural parameter η are related as $p = \frac{e^\eta}{1+e^\eta}$. Using this, and our general formulas from Theorem 5.3, we can rewrite the mean, variance, skewness, and kurtosis of X as

$$E(X) = np;\ \mathrm{Var}(X) = np(1-p);\ \beta_p = \frac{1-2p}{\sqrt{np(1-p)}};\ \gamma_p = \frac{\frac{1}{p(1-p)} - 6}{n}.$$

For completeness, it is useful to have the mean and the variance formula in an original parametrization, and they are stated below. The proof follows from an application of Theorem 5.3 and the chain rule.

Theorem 3.6. Let $\{P_\theta, \theta \in \Theta\}$ be a family of distributions in the one parameter Exponential family with density (pmf)

$$f(x\,|\theta) = e^{\eta(\theta)T(x)-\psi(\theta)}h(x).$$

Then, at any θ at which $\eta'(\theta) \neq 0$,

$$E_\theta[T(X)] = \frac{\psi'(\theta)}{\eta'(\theta)};\ \mathrm{Var}_\theta(T(X)) = \frac{\psi''(\theta)}{[\eta'(\theta)]^2} - \frac{\psi'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3}.$$

3.3.3 Closure Properties

The Exponential family satisfies a number of important closure properties. For instance, if a d-dimensional random vector $X = (X_1, \cdots, X_d)$ has a distribution in the Exponential family, then the conditional distribution of any subvector given the rest is also in the Exponential family. There are a number of such closure properties, of which we will discuss only four.

First, if $X = (X_1, \cdots, X_d)$ has a distribution in the Exponential family, then the natural sufficient statistic $T(X)$ also has a distribution in the Exponential family. Verification of this in the greatest generality cannot be done without using measure theory. However, we can easily demonstrate this in some particular cases. Consider the continuous case with $d = 1$ and suppose $T(X)$ is a differentiable one-to-one function of X. Then, by the Jacobian formula (see Chapter 3), $T(X)$ has the density

$$f_T(t\,|\eta) = e^{\eta t-\psi(\eta)}\frac{h(T^{-1}(t))}{|T'(T^{-1}(t))|}.$$

This is once again in the one parameter Exponential family form, with the natural sufficient statistic as T itself, and the ψ function unchanged. The h function has changed to a new function $h^*(t) = \frac{h(T^{-1}(t))}{|T'(T^{-1}(t))|}$.

Similarly, in the discrete case, the pmf of $T(X)$ will be given by

$$P_\eta(T(X) = t) = \sum_{x:\,T(x)=t} e^{\eta T(x)-\psi(\eta)}h(x) = e^{\eta t-\psi(\eta)}h^*(t),$$

where $h^*(t) = \sum_{x:\,T(x)=t} h(x)$.

Next, suppose $X = (X_1, \cdots, X_d)$ has a density (pmf) $f(x\,|\eta)$ in the Exponential family and

$Y_1, Y_2, \cdots, Y_n$ are n iid observations from this density $f(x\,|\eta)$. Note that each individual Y_i is a d-dimensional vector. The joint density of $Y = (Y_1, Y_2, \cdots, Y_n)$ is

$$f(y\,|\eta) = \prod_{i=1}^{n} f(y_i\,|\eta) = \prod_{i=1}^{n} e^{\eta T(y_i) - \psi(\eta)} h(y_i)$$

$$= e^{\eta \sum_{i=1}^{n} T(y_i) - n\psi(\eta)} \prod_{i=1}^{n} h(y_i).$$

We recognize this to be in the one parameter Exponential family form again, with the natural sufficient statistic as $\sum_{i=1}^{n} T(Y_i)$, the new ψ function as $n\psi$, and the new h function as $\prod_{i=1}^{n} h(y_i)$.

The joint density $\prod_{i=1}^{n} f(y_i\,|\eta)$ is known as *the likelihood function* in statistics. So, likelihood functions obtained from an iid sample from a distribution in the one parameter Exponential family are also members of the one parameter Exponential family.

The closure properties outlined in the above are formally stated in the next theorem.

Theorem 3.7. Suppose $X = (X_1, \cdots, X_d)$ has a distribution belonging to the one parameter Exponential family with the natural sufficient statistic $T(X)$.

(a) $T = T(X)$ also has a distribution belonging to the one parameter Exponential family.

(b) Let $Y = AX + u$ be a nonsingular linear transformation of X. Then Y also has a distribution belonging to the one parameter Exponential family.

(c) Let $\mathcal{I}_0$ be any proper subset of $\mathcal{I} = \{1, 2, \cdots, d\}$. Then the joint conditional distribution of $X_i, i \in \mathcal{I}_0$ given $X_j, j \in \mathcal{I} - \mathcal{I}_0$ also belongs to the one parameter Exponential family.

(d) For given $n \geq 1$, suppose $Y_1, \cdots, Y_n$ are iid with the same distribution as X. Then the joint distribution of $(Y_1, \cdots, Y_n)$ also belongs to the one parameter Exponential family.

3.3.4 Curved Exponential Family

There are some important examples in which the density (pmf) has the basic Exponential family form $f(x\,|\theta) = e^{\sum_{i=1}^{k} \eta_i(\theta) T_i(X) - \psi(\theta)} h(x)$, but the assumption that the dimensions of Θ, and that of the range space of $(\eta_1(\theta), \cdots, \eta_k(\theta))$ are the same is violated. More precisely, the dimension of Θ is some positive integer q strictly less than k. Let us start with an example.

Example 3.14. Suppose $X \sim N(\mu, \mu^2), \mu \neq 0$. Writing $\mu = \theta$, the density of X is

$$f(x\,|\theta) = \frac{1}{\sqrt{2\pi}|\theta|} e^{-\frac{1}{2\theta^2}(x-\theta)^2} I_{x \in \mathcal{R}}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\theta^2} + \frac{x}{\theta} - \frac{1}{2} - \log|\theta|} I_{x \in \mathcal{R}}.$$

Writing $\eta_1(\theta) = -\frac{1}{2\theta^2}, \eta_2(\theta) = \frac{1}{\theta}, T_1(x) = x^2, T_2(x) = x, \psi(\theta) = \frac{1}{2} + \log|\theta|$, and $h(x) = \frac{1}{\sqrt{2\pi}} I_{x \in \mathcal{R}}$, this is in the form $f(x\,|\theta) = e^{\sum_{i=1}^k \eta_i(\theta)T_i(x) - \psi(\theta)} h(x)$, with $k = 2$, although $\theta \in \mathcal{R}$, which is only one dimensional. The two functions $\eta_1(\theta) = -\frac{1}{2\theta^2}$ and $\eta_2(\theta) = \frac{1}{\theta}$ are related to each other by the identity $\eta_1 = -\frac{\eta_2^2}{2}$, so that a plot of (η_1, η_2) in the plane would be a curve, not a straight line. Distributions of this kind go by the name of *curved Exponential family.* The dimension of the natural sufficient statistic is more than the dimension of Θ for such distributions.

Definition 3.7. Let $X = (X_1, \cdots, X_d)$ have a distribution $P_\theta, \theta \in \Theta \subseteq \mathcal{R}^q$. Suppose P_θ has a density (pmf) of the form

$$f(x\,|\theta) = e^{\sum_{i=1}^k \eta_i(\theta)T_i(x) - \psi(\theta)} h(x),$$

where $k > q$. Then, the family $\{P_\theta, \theta \in \Theta\}$ is called a *curved Exponential family.*

Example 3.15. (A Specific Bivariate Normal). Suppose $X = (X_1, X_2)$ has a bivariate normal distribution with zero means, standard deviations equal to one, and a correlation parameter $\rho, -1 < \rho < 1$. The density of X is

$$f(x\,|\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[x_1^2 + x_2^2 - 2\rho x_1 x_2\right]} I_{x_1, x_2 \in \mathcal{R}}$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x_1^2+x_2^2}{2(1-\rho^2)} + \frac{\rho}{1-\rho^2} x_1 x_2} I_{x_1, x_2 \in \mathcal{R}}.$$

Therefore, here we have a curved Exponential family with $q = 1, k = 2, \eta_1(\rho) = -\frac{1}{2(1-\rho^2)}, \eta_2(\rho) = \frac{\rho}{1-\rho^2}, T_1(x) = x_1^2 + x_2^2, T_2(x) = x_1 x_2, \psi(\rho) = \frac{1}{2}\log(1-\rho^2)$, and $h(x) = \frac{1}{2\pi} I_{x_1, x_2 \in \mathcal{R}}$.

Example 3.16. (Poissons with Random Covariates). Suppose given $Z_i = z_i, i = 1, 2, \cdots, n, X_i$ are independent $Poi(\lambda z_i)$ variables, and $Z_1, Z_2, \cdots, Z_n$ have some joint pmf $p(z_1, z_2, \cdots, z_n)$. It is implicitly assumed that each $Z_i > 0$ with probability one. Then, the joint pmf of $(X_1, X_2, \cdots, X_n, Z_1, Z_2, \cdots, Z_n)$ is

$$f(x_1, \cdots, x_n, z_1, \cdots, z_n\,|\lambda) = \prod_{i=1}^n \frac{e^{-\lambda z_i}(\lambda z_i)^{x_i}}{x_i!} p(z_1, z_2, \cdots, z_n) I_{x_1, \cdots, x_n \in \mathcal{N}_0} I_{z_1, z_2, \cdots, z_n \in \mathcal{N}_1}$$

$$= e^{-\lambda \sum_{i=1}^n z_i + (\sum_{i=1}^n x_i)\log\lambda} \prod_{i=1}^n \frac{z_i^{x_i}}{x_i!} p(z_1, z_2, \cdots, z_n) I_{x_1, \cdots, x_n \in \mathcal{N}_0} I_{z_1, z_2, \cdots, z_n \in \mathcal{N}_1},$$

where $\mathcal{N}_0$ is the set of nonnegative integers, and $\mathcal{N}_1$ is the set of positive integers.
This is in the curved Exponential family with

$$q = 1, k = 2, \eta_1(\lambda) = -\lambda, \eta_2(\lambda) = \log\lambda, T_1(x, z) = \sum_{i=1}^n z_i, T_2(x, z) = \sum_{i=1}^n x_i,$$

and

$$h(x,z) = \prod_{i=1}^{n} \frac{z_i^{x_i}}{x_i!} p(z_1, z_2, \cdots, z_n) I_{x_1,\cdots,x_n \in \mathcal{N}_0} I_{z_1,z_2,\cdots,z_n \in \mathcal{N}_1}.$$

If we consider the covariates as fixed, the joint distribution of $(X_1, X_2, \cdots, X_n)$ becomes a regular one parameter Exponential family.

3.4 Exercises

Exercise 3.1. (Poisson Distribution). Show that the Poisson distribution belongs to the one parameter Exponential family if $\lambda > 0$. Write it in the canonical form and by using the mean parametrization.

Exercise 3.2. (Negative Binomial Distribution). Show that the negative binomial distribution with parameters r and p belongs to the one parameter Exponential family if r is considered fixed and $0 < p < 1$. Write it in the canonical form and by using the mean parametrization.

Exercise 3.3. (Generalized Negative Binomial Distribution). Show that the generalized negative binomial distribution with the pmf $f(x\,|p) = \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!} p^{\alpha}(1-p)^x, x = 0, 1, 2, \cdots$ belongs to the one parameter Exponential family if $\alpha > 0$ is considered fixed and $0 < p < 1$.

Exercise 3.4. (Generalized Negative Binomial Distribution). Show that the two parameter generalized negative binomial distribution with the pmf $f(x\,|\alpha, p) = \frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!} p^{\alpha}(1-p)^x, x = 0, 1, 2, \cdots$ does not belong to the two parameter Exponential family.

Exercise 3.5. (Hardy-Weinberg Law). Suppose genotypes at a single locus with two alleles are present in a population according to the relative frequencies $p^2, 2pq$, and q^2, where $q = 1-p$, and p is the relative frequency of the dominant allele. Show that the joint distribution of the frequencies of the three genotypes in a random sample of n individuals from this population belongs to a one parameter Exponential family if $0 < p < 1$. Write it in the canonical form and by using the mean parametrization.

Exercise 3.6. (Beta Distribution). Show that the two parameter Beta distribution belongs to the two parameter Exponential family if the parameters $\alpha, \beta > 0$. Write it in the canonical form and by using the mean parametrization.

Exercise 3.7. (Beta Distribution). Show that symmetric Beta distributions belong to the one parameter Exponential family if the single parameter $\alpha > 0$.

Exercise 3.8. (Normal with Equal Mean and Variance). Show that the $N(\theta, \theta)$ distribution belongs to the one parameter Exponential family if $\theta > 0$. Write it in the canonical form and by using the mean parametrization.

Exercise 3.9. (Truncated Poisson). The number of fires reported in a week to a city fire station is Poisson with some mean λ. The city station is supposed to report the number each week to the central state office. But they do not bother to report it if their number of reports is less than 3. Suppose you are employed at the state central office. Model the problem, and prove or disprove that the pmf you chose is in an Exponential family.

Exercise 3.10. (Binomial with Both Parameters Unknown). Suppose $X_1, \cdots, X_k \overset{iid}{\sim} Bin((N,p), N \geq 1, 0 < p < 1$. Is this a two parameter Exponential family?

Exercise 3.11. (Beta-Binomial). Suppose given $p, X \sim Bin(n,p)$, where n is known, and that $p \sim Beta(\alpha, \beta), \alpha, \beta$ considered to be unknown. Is the marginal pmf of X an Exponential family?

Exercise 3.12. (Identifiability of the Distribution). Show that distributions in the nonsingular canonical one parameter Exponential family are identifiable, i.e., $P_{\eta_1} = P_{\eta_2}$ only if $\eta_1 = \eta_2$.

Exercise 3.13. (Infinite Differentiability of Mean Functionals). Suppose $P_\theta, \theta \in \Theta$ is a one parameter Exponential family and $\phi(x)$ is a general function. Show that at any $\theta \in \Theta^0$ at which $E_\theta[|\phi(X)|] < \infty, \mu_\phi(\theta) = E_\theta[\phi(X)]$ is infinitely differentiable, and can be differentiated any number of times inside the integral (sum).

Exercise 3.14. (Poisson Skewness and Kurtosis). Find the skewness and kurtosis of a Poisson distribution by using Theorem 5.3.

Exercise 3.15. (Gamma Skewness and Kurtosis). Find the skewness and kurtosis of a Gamma distribution, considering α as fixed, by using Theorem 5.3.

Exercise 3.16. (Multinomial Covariances). Calculate the covariances in a multinomial distribution by using Theorem 5.7.

Exercise 3.17. (Distributions with Zero Skewness). Show the remarkable result that the only distributions in a canonical one parameter Exponential family such that the natural sufficient statistic has a zero skewness are the normal distributions with a fixed variance.

Exercise 3.18. (Dirichlet Distribution). Show that the Dirichlet distribution with parameter vector $\alpha = (\alpha_1, \cdots, \alpha_{n+1}), \alpha_i > 0$ for all i, is an $(n+1)$-parameter Exponential family.

Exercise 3.19. (Normal Linear Model). Suppose given an $n \times p$ nonrandom matrix X, a parameter vector $\beta \in \mathcal{R}^p$, and a variance parameter $\sigma^2 > 0$, $Y = (Y_1, Y_2, \cdots, Y_n) \sim N_n(X\beta, \sigma^2 I_n)$, where I_n is the $n \times n$ identity matrix. Show that the distribution of Y belongs to a full rank multiparameter Exponential family.

Exercise 3.20. (A Special Bivariate Normal). Consider the bivariate normal density with equal means μ, μ, equal variances σ^2, σ^2 and correlation ρ; μ, σ, ρ are considered unknown parameters. Is this a regular Exponential family, or a curved Exponential family, or neither?

Exercise 3.21. (Normal with an Integer Mean). Suppose $X \sim N(\mu, 1)$, where μ is known to be an integer. Is this a regular one parameter Exponential family?

Exercise 3.22. (Mixtures of Normal). Show that the mixture distribution $.5N(\mu, 1) + .5N(\mu, 100)$ does not belong to the one parameter Exponential family. Generalize this result to more general mixtures of normal distributions.

Exercise 3.23. (Double Exponential Distribution). (a) Show that the double exponential distribution with a known σ value and an unknown mean does not belong to the one parameter Exponential family, but the double exponential distribution with a known mean and an unknown σ belongs to the one parameter Exponential family.
(b) Show that the two parameter double exponential distribution does not belong to the two parameter Exponential family.

Exercise 3.24. (A Curved Exponential Family). Suppose $(X_1, X_2, \cdots, X_n)$ are jointly multivariate normal with general means μ_i, variances all one, and a common pairwise correlation ρ. Show that the distribution of $(X_1, X_2, \cdots, X_n)$ is a curved Exponential family.

Exercise 3.25. (Poissons with Covariates). Suppose $X_1, X_2, \cdots, X_n$ are independent Poissons with $E(X_i) = \lambda e^{\beta z_i}, \lambda > 0, -\infty < \beta < \infty$. The covariates $z_1, z_2, \cdots, z_n$ are considered fixed. Show that the distribution of $(X_1, X_2, \cdots, X_n)$ is a curved Exponential family.

Exercise 3.26. (Another Curved Exponential Family). Suppose $X \sim Bin(n, p), Y \sim Bin(m, p^2)$, and that X, Y are independent. Show that the joint distribution of (X, Y) is a curved Exponential family.

3.5 References

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

Bickel, P. J. and Doksum, K. (2006). Mathematical Statistics, Basic Ideas and Selected Topics, Vol I, Prentice Hall, Saddle River, NJ.

Brown, L. D. (1986). Fundamentals of Statistical Exponential Families, IMS, Lecture Notes and Monographs Series, Hayward, CA.

Lehmann, E. L. and Casella, G. (1998). Theory of Point Estimation, Springer, New York.

LeTac, G. (1992). Lectures on natural exponential families and their variance functions, Monogr. Mat., 50, Mat. Pura. Aplic. Rio.

Liese, F. and Miescke, K. (2008). Statistical Decision Theory, Springer, New York.