Lecture Notes for STAT-F404, author: Marc Hallin[1]

4 Exponential Families

4.1 Definitions

Exponential families of distributions are parametric dominated families in which the logarithm of probability densities take a simple bilinear form (bilinear in the parameter and in a statistic). As a consequence of that special form, a number of statistical problems, in such families, are well-posed, and can be solved. In particular, the results on point estimation developed in the previous chapter straightforwardly apply. Many traditional families of distributions—binomial, multinomial, Poisson, negative binomial, normal, gamma, chi-square, beta, Dirichlet, Wishart, and many others—are exponential families. Note, however, that the uniform, logistic, Cauchy, or Student (with fixed degrees of freedom) location-scale families are not exponential; the double-exponential (or Laplace) family is exponential for scale at fixed location, but not for location at fixed scale.

A statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ (or the family $\mathcal{P}$ itself) is called *exponential* if $\mathcal{P}$ is a parametric family $\mathcal{P} = \left\{\mathrm{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \Theta \subseteq \mathbb{R}^K\right\}$, dominated by some σ-finite measure μ, with corresponding densities $f_{\boldsymbol{\theta}} := \frac{\mathrm{dP}_{\boldsymbol{\theta}}}{\mathrm{d}\mu}$ taking a value, at $\mathbf{x} \in \mathcal{X}$, of the form

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = C(\boldsymbol{\theta}) h(\mathbf{x}) \exp\left(\sum_{k=1}^{K} \theta_k T_k(\mathbf{x})\right), \tag{4.1}$$

where $\mathbf{T} := (T_1, \ldots, T_K)$ is a statistic with values in $(\mathbb{R}^K, \mathcal{B}^K)$ and where Θ is a subset of

$$\Theta_0 := \left\{\mathbf{w} \in \mathbb{R}^K : \int_{\mathcal{X}} h(\mathbf{x}) \exp\left(\sum_{k=1}^{K} w_k T_k(\mathbf{x})\right) \mathrm{d}\mu(\mathbf{x}) < \infty\right\}. \tag{4.2}$$

The statistic $\mathbf{T}$ is called a *natural* or *privileged statistic*[2], $\boldsymbol{\theta}$ a *natural parameter*, and Θ_0 the *natural parameter space* of the family.

[1] With slight modifications by Davy Paindaveine and Thomas Verdebout.

[2] It readily follows from the factorization criterion that this statistic $\mathbf{T}$ is sufficient, so that the terminology "sufficient statistic" is often used; this is, however, slightly misleading, since any bijective transformation of $\mathbf{T}$ is sufficient, but would not qualify as a "natural" or "privileged" statistic.

Remark 1: The function $h(\mathbf{x})$ plays no structural role, and is related to the choice of a dominating measure μ. With the measure defined by

$$\nu(B) := \int_B h(\mathbf{x}) \mathrm{d}\mu(\mathbf{x}),$$

we have $\mathrm{P}_{\boldsymbol{\theta}} << \nu << \mu$, and the densities with respect to this alternative dominating measure take the form

$$\frac{\mathrm{dP}_{\boldsymbol{\theta}}}{\mathrm{d}\nu} = \frac{\mathrm{dP}_{\boldsymbol{\theta}}}{\mathrm{d}\mu} \Big/ \frac{\mathrm{d}\nu}{\mathrm{d}\mu} = f_{\boldsymbol{\theta}}(\mathbf{x})/h(\mathbf{x}) = C(\boldsymbol{\theta}) \exp\left(\sum_{k=1}^{K} \theta_k T_k(\mathbf{x})\right).$$

This shows that the function h can thus be "absorbed into μ".

Remark 2: The density $f_{\boldsymbol{\theta}}$ also can be written as

$$f_{\boldsymbol{\theta}} = C(\boldsymbol{\theta}) h(\mathbf{x}) \exp(\boldsymbol{\theta}' \mathbf{T}(\mathbf{x})) = C(\boldsymbol{\theta}) h(\mathbf{x}) \exp(\{\mathbf{A}'\boldsymbol{\theta}\}'\{\mathbf{A}^{-1}\mathbf{T}(\mathbf{x})\}),$$

where $\mathbf{A}$ is an arbitrary invertible $K \times K$ matrix: thus, $\boldsymbol{\theta}$ and $\mathbf{T}$ are not uniquely identified.

Remark 3: The natural parameter space Θ_0 is the largest subset of $\mathbb{R}^K$ such that the exponential can be normed into a probability density, and $C(\boldsymbol{\theta})$ is then the inverse of the integral in (4.2). That natural parameter space is a *convex* subset of $\mathbb{R}^K$. Indeed, if $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$ belong to Θ_0, then, for any convex linear combination $\boldsymbol{\theta} = \alpha \boldsymbol{\theta}' + (1-\alpha) \boldsymbol{\theta}''$, $\alpha \in (0, 1)$, Hölder's inequality yields

$$\int h(\mathbf{x}) \exp\left(\sum_{k=1}^{K} \{\alpha \boldsymbol{\theta}' + (1-\alpha) \boldsymbol{\theta}''\} T_k(\mathbf{x})\right) d\mu(\mathbf{x})$$
$$\leq \left\{\int h(\mathbf{x}) \exp\left(\sum_{k=1}^{K} \theta_k' T_k(\mathbf{x})\right) d\mu(\mathbf{x})\right\}^{\alpha} \left\{\int h(\mathbf{x}) \exp\left(\sum_{k=1}^{K} \theta_k'' T_k(\mathbf{x})\right) d\mu(\mathbf{x})\right\}^{1-\alpha} < \infty,$$

which shows that $\boldsymbol{\theta}$ also belongs to the natural parameter space Θ_0.

Remark 4: If $\mathcal{P}$ is an exponential family with natural parameter $\boldsymbol{\theta}$ and privileged statistic $\mathbf{T}$, then the induced family $\mathcal{P}^{\mathbf{T}}$ is also exponential with natural parameter $\boldsymbol{\theta}$ and privileged statistic $\mathbf{T}$

(here the identity statistic, since one observes a value $\mathbf{t}$ of $\mathbf{T}$ in the induced model), meaning that

$$f_{\boldsymbol{\theta}}^{\mathbf{T}}(\mathbf{t}) := \frac{\mathrm{dP}_{\boldsymbol{\theta}}^{\mathbf{T}}}{\mathrm{d}\mu^{\mathbf{T}}}(\mathbf{t}) = C(\boldsymbol{\theta}) \exp\left(\sum_{k=1}^{K} \boldsymbol{\theta}_k t_k\right), \qquad \mathbf{t} \in \mathbb{R}^K$$

(here, we assumed that μ was chosen to absorb the factor $h(\mathbf{x})$).

<u>Remark 5</u>: Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, where $\mathbf{X}_i$, $i = 1, \ldots, n$, are identically and independently distributed with exponential density

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = C(\boldsymbol{\theta}) h(\mathbf{x}) \exp\left(\sum_{k=1}^{K} \theta_k T_k(\mathbf{x})\right).$$

Then, $\mathbf{X}$ also has an exponential density, of the form

$$f_{\boldsymbol{\theta}}^{\mathbf{X}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = (C(\boldsymbol{\theta}))^n \left(\prod_{i=1}^{n} h(\mathbf{x}_i)\right) \exp\left(\sum_{k=1}^{K} \theta_k \left\{\sum_{i=1}^{n} T_k(\mathbf{x}_i)\right\}\right),$$

which involves the same natural parameter $\boldsymbol{\theta}$ as $\mathbf{X}_1$, and the privileged statistic $\sum_{i=1}^{n} \mathbf{T}(\mathbf{X}_i)$. We refer to Section 4.4 for an important consequence of this fact.

Definition 1. *We say that the exponential family $\mathcal{P}$ with densities $f_{\boldsymbol{\theta}}$ given in (4.1) has* full rank K *if its parameter space Θ contains at least a hypercube with strictly positive Lebesgue measure in $(\mathbb{R}^K, \mathcal{B}^K)$—equivalently, if the interior of Θ is not void.*

Two reasons may cause an exponential family with K-dimensional natural parameter to have rank less than K:

- Θ is a linear manifold of $\mathbb{R}^K$: then, the parameter and the privileged statistic can be redefined via linear transformations of the original ones. Being linear, those transformations do not affect the bilinear nature of the exponential part of $f_{\boldsymbol{\theta}}$, yielding a new form for the densities, as an exponential family of full rank $K' < K$ (see Example 6 below);

- Θ is a *nonlinear* manifold of $\mathbb{R}^K$: then the family is called a *curved exponential family* and loses many of the appealing properties of (full rank) exponential families. That case will not be considered here.

4.2 Examples

As already mentioned, many common families of distributions are exponential families. We provide some examples.

Example 1: the *Bernoulli family*. Bernoulli distributions over $(\mathbb{R}, \mathcal{B})$, parametrized by $p \in (0, 1)$, have densities with respect to the counting measure of $\{0, 1\}$ given by

$$f_p(x) = p^x (1-p)^{(1-x)}.$$

We have

$$\begin{aligned} f_p(x) &= \exp[\log f_p(x)] = \exp[x \log p + (1-x)\log(1-p)] \\ &= \exp[\log(1-p)] \exp[x(\log p - \log(1-p))] \\ &= (1-p)\exp[x \log(p/(1-p))]. \end{aligned}$$

This has the exponential form (4.1), with $h(x) = 1$, natural parameter $\theta = \log(p/(1-p))$, privileged statistic $T(x) = x$, and $C(\theta) = 1 - p = 1/(1 + \exp(\theta))$.

Example 2: the *Binomial family*. For given $n \in \mathbb{N}_0$, Binomial distributions over $(\mathbb{R}, \mathcal{B})$, parametrized by $p \in (0, 1)$, have densities with respect to the counting measure of $\{0, 1, \ldots, n\}$ given by

$$f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

The same computation as above yields

$$f_p(x) = \exp[\log(f_p(x))] = (1-p)^n \binom{n}{x} \exp[x \log(p/(1-p))],$$

from where it is clear that the binomial family is exponential, with (recall that n is a constant here, not a parameter) $h(x) = \binom{n}{x}$, natural parameter $\theta = \log(p/(1-p))$, privileged statistic $T(x) = x$, and $C(\theta) = (1-p)^n = 1/(1 + \exp(\theta))^n$.

Example 3: the *Poisson family*. Poisson distributions over $(\mathbb{R}, \mathcal{B})$, with parameter $\lambda \in \mathbb{R}_0^+$, have densities with respect to the counting measure of $\mathbb{N}$ given by

$$f_\lambda(x) = \exp(-\lambda)\frac{\lambda^x}{x!}.$$

One readily obtains

$$f_\lambda(x) = \exp(-\lambda)\frac{1}{x!}\exp(x \log \lambda).$$

Hence, the Poisson family is exponential, with $h(x) = 1/x!$, natural parameter $\theta = \log \lambda$, privileged statistic $T(x) = x$, and $C(\theta) = \exp(-\lambda) = \exp(-\exp(\theta))$.

Example 4: the *Gaussian family*. Gaussian distributions on $(\mathbb{R}, \mathcal{B})$, parametrized by $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_0^+$, have densities with respect to the Lebesgue measure

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

The canonical exponential family form (4.1) is easily obtained as

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2 - 2\mu x + \mu^2}{2\sigma^2}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right).$$

The Gaussian family is thus exponential, with $h(x) = 1$, natural parameter $\boldsymbol{\theta} = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})'$, privileged statistic $\mathbf{T}(x) = (x, x^2)'$, and

$$C(\boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) = \sqrt{\frac{-\theta_2}{\pi}} \exp\left(\frac{\theta_1^2}{4\theta_2}\right).$$

Example 5: the *Lognormal family*. Lognormal distributions, like the Gaussian ones, have densities, parametrized by $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$, of the form

$$f_{\mu,\sigma^2}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{((\log x) - \mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}_0^+,$$

with respect to the Lebesgue measure on $(\mathbb{R}_0^+, \mathcal{B} \cap \mathbb{R}_0^+)$. One easily obtains

$$f_{\mu,\sigma^2}(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(\frac{\mu}{\sigma^2}\log x - \frac{1}{2\sigma^2}(\log x)^2\right);$$

the Lognormal family is thus exponential, with $h(x) = 1/x$, natural parameter $\boldsymbol{\theta} = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})'$, privileged statistic $\mathbf{T}(x) = (\log x, (\log x)^2)'$, and

$$C(\boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) = \sqrt{\frac{-\theta_2}{\pi}} \exp\left(\frac{\theta_1^2}{4\theta_2}\right).$$

Example 6: the *Multinomial family*. For given exponent $n \in \mathbb{N}_0$, multinomial distributions are parametrized by a probability vector $\mathbf{p} = (p_1, \ldots, p_K)' \in [0,1]^K$ such that $\sum_{k=1}^K p_k = 1$. The probability density, over $(\mathbb{R}^K, \mathbb{B}^K)$ and with respect to the counting measure of points $\mathbf{x} \in \mathbb{R}^K$ with integer-valued coordinates $x_1, \ldots, x_K$ such that $\sum_{k=1}^K x_k = n$, is

$$f_{\mathbf{p}}(\mathbf{x}) = \frac{n!}{x_1! \ldots x_K!} p_1^{x_1} \ldots p_K^{x_K} = \frac{n!}{x_1! \ldots x_K!} \exp\left(\sum_{k=1}^K x_k \log p_k\right).$$

The multinomial family with exponent n and parameter $\mathbf{p}$ is thus also an exponential family, with $h(\mathbf{x}) = n!/(x_1! \ldots x_K!)$, natural parameter $\boldsymbol{\theta} = (\log p_1, \ldots, \log p_K)'$, privileged statistic $\mathbf{T}(\mathbf{x}) = \mathbf{x}$, and $C(\boldsymbol{\theta}) = 1$. But, since $\sum_{k=1}^K p_k = 1$, the parameter space of that family is a $(K-1)$-dimensional linear manifold of $\mathbb{R}^K$, and the family does not have full rank K. Replacing p_K with $1 - \sum_{k=1}^{K-1} p_k$ and x_K with $n - \sum_{k=1}^{K-1} x_k$, the same densities rewrite

$$\begin{aligned} f_{p_1 \ldots p_{K-1}}(x_1, \ldots, x_{K-1}) \; = \; & \frac{n!}{x_1! \ldots x_{K-1}!(n - \sum_{k=1}^{K-1} x_k)!} \left(1 - \sum_{k=1}^{K-1} p_k\right)^n \\ & \times \exp\left(x_1 \log\left(\frac{p_1}{1 - \sum_{k=1}^{K-1} p_k}\right) + \ldots + x_{K-1} \log\left(\frac{p_{K-1}}{1 - \sum_{k=1}^{K-1} p_k}\right)\right). \end{aligned}$$

It follows that the multinomial family actually is exponential with full rank $K-1$ instead of K,

$$h(x_1, \ldots, x_{K-1}) = \frac{n!}{x_1! \ldots x_{K-1}!(n - \sum_{k=1}^{K-1} x_k)!},$$

natural parameter $\boldsymbol{\theta} = \big(\log(p_1/(1-\sum_{k=1}^{K-1} p_k)), \ldots, \log(p_{K-1}/(1-\sum_{k=1}^{K-1} p_k))\big)'$, privileged statistic $\mathbf{T}(\mathbf{x}) = (x_1, \ldots, x_{K-1})'$, and $C(\boldsymbol{\theta}) = \big(1 - \sum_{k=1}^{K-1} p_k\big)^n = \big(1 + \sum_{k=1}^{K-1} \exp(\theta_k)\big)^{-n}$.

4.3 Sufficiency and completeness in exponential families

It readily follows from the factorization criterion that the privileged statistic $\mathbf{T}$ of an exponential family is *sufficient*. The following proposition shows that, if the family has full rank, then $\mathbf{T}$ is also minimal sufficient.

Proposition 1. *If an exponential family $\mathcal{P}$ with privileged statistic $\mathbf{T}$ has full rank K, then $\mathbf{T}$ is* minimal sufficient.

Proof. If the model has full rank K, then Θ contains at least a $(K+1)$-tuple $\big(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}\big)$ determining a hyperrectangle with positive volume, that is, such that

$$\det \begin{pmatrix} \theta_1^{(1)} - \theta_1^{(0)} & \theta_1^{(2)} - \theta_1^{(0)} & \cdots & \theta_1^{(K)} - \theta_1^{(0)} \\ \vdots & \vdots & & \vdots \\ \theta_K^{(1)} - \theta_K^{(0)} & \theta_K^{(2)} - \theta_K^{(0)} & \cdots & \theta_K^{(K)} - \theta_K^{(0)} \end{pmatrix} \neq 0 \, . \tag{4.3}$$

Consider the subfamily $\mathcal{P}_0 = \{\mathrm{P}_{\boldsymbol{\theta}^{(0)}}, \ldots, \mathrm{P}_{\boldsymbol{\theta}^{(K)}}\}$, and let us apply the method described in Chapter 2 for constructing minimal sufficient statistics. We know that it suffices to show that $\mathbf{T}$ is minimal sufficient for $\mathcal{P}_0$ (it indeed follows from the Neyman-Fisher factorization criterion that $\mathbf{T}$ is sufficient for $\mathcal{P}$). Since $\mathrm{P}_{\boldsymbol{\theta}^{(i)}} << \mathrm{P}_{\boldsymbol{\theta}^{(0)}}$ for $i = 1, \ldots, K$,

$$\mathbf{S} := \left(\frac{\mathrm{dP}_{\boldsymbol{\theta}^{(1)}}}{\mathrm{dP}_{\boldsymbol{\theta}^{(0)}}}, \ldots, \frac{\mathrm{dP}_{\boldsymbol{\theta}^{(K)}}}{\mathrm{dP}_{\boldsymbol{\theta}^{(0)}}} \right)'$$

is a minimal sufficient statistic for $\mathcal{P}_0$. Consequently, the statistic $\tilde{\mathbf{S}} := (\tilde{S}_1, \ldots, \tilde{S}_K)'$, with

$$\tilde{S}_i := \log \left(\frac{C(\boldsymbol{\theta}^{(0)})}{C(\boldsymbol{\theta}^{(i)})} \frac{\mathrm{dP}_{\boldsymbol{\theta}^{(i)}}}{\mathrm{dP}_{\boldsymbol{\theta}^{(0)}}} \right) = \log \left(\frac{C(\boldsymbol{\theta}^{(0)}) f_{\boldsymbol{\theta}^{(i)}}}{C(\boldsymbol{\theta}^{(i)}) f_{\boldsymbol{\theta}^{(0)}}} \right) = \sum_{k=1}^{K} \big(\theta_k^{(i)} - \theta_k^{(0)}\big) T_k, \qquad i = 1, \ldots, K,$$

is also minimal sufficient for $\mathcal{P}_0$ (since it is a one-to-one function of $\mathbf{S}$). Under matrix form, we have

$$\tilde{\mathbf{S}} = \begin{pmatrix} \theta_1^{(1)} - \theta_1^{(0)} & \cdots & \theta_K^{(1)} - \theta_K^{(0)} \\ \theta_1^{(2)} - \theta_1^{(0)} & \cdots & \theta_K^{(2)} - \theta_K^{(0)} \\ \vdots & & \vdots \\ \theta_1^{(K)} - \theta_1^{(0)} & \cdots & \theta_K^{(K)} - \theta_K^{(0)} \end{pmatrix} \mathbf{T},$$

which, in view of (4.3), shows that $\mathbf{T}$ is a one-to-one function of $\tilde{\mathbf{S}}$. As a one-to-one function of a minimal sufficient statistic for $\mathcal{P}_0$, the privileged statistic is thus also minimal sufficient for $\mathcal{P}_0$. $\square$

Under the same conditions as in the previous result, $\mathbf{T}$ is also complete.

Proposition 2. *If an exponential family $\mathcal{P}$ with privileged statistic $\mathbf{T}$ has full rank K, then $\mathbf{T}$ is* complete.

Proof. Let the function $\ell : \mathcal{T} \to \mathbb{R}$ be such that $\mathrm{E}_{\boldsymbol{\theta}}[\ell(\mathbf{T})] = 0$ for any $\boldsymbol{\theta} \in \Theta$. Working with the dominating measure ν defined in Remark 1, this implies that

$$C(\boldsymbol{\theta}) \int_{\mathcal{T}} \ell(\mathbf{t}) \exp(\boldsymbol{\theta}'\mathbf{t}) d\nu^{\mathbf{T}}(\mathbf{t}) = C(\boldsymbol{\theta}) \int_{\mathcal{X}} \ell(\mathbf{T}(\mathbf{x})) \exp(\boldsymbol{\theta}'\mathbf{T}(\mathbf{x})) d\nu(\mathbf{x}) = 0$$

for any $\boldsymbol{\theta} \in \Theta$. Writing $\ell = \ell^+ - \ell^-$, where both ℓ^+ and ℓ^- take their values in $\mathbb{R}^+$ (as we did when defining conditional expectations in Chapter 2), we thus have

$$\int_{\mathcal{T}} \ell^+(\mathbf{t}) \exp(\boldsymbol{\theta}'\mathbf{t}) d\nu^{\mathbf{T}}(\mathbf{t}) = \int_{\mathcal{T}} \ell^-(\mathbf{t}) \exp(\boldsymbol{\theta}'\mathbf{t}) d\nu^{\mathbf{T}}(\mathbf{t}) \tag{4.4}$$

for any $\boldsymbol{\theta} \in \Theta$. Fix $\boldsymbol{\theta}_0$ in the interior of Θ, and put

$$C := \int_{\mathcal{T}} \ell^+(\mathbf{t}) \exp(\boldsymbol{\theta}_0'\mathbf{t}) d\nu^{\mathbf{T}}(\mathbf{t}) = \int_{\mathcal{T}} \ell^-(\mathbf{t}) \exp(\boldsymbol{\theta}_0'\mathbf{t}) d\nu^{\mathbf{T}}(\mathbf{t}) \tag{4.5}$$

Obviously, $C \geq 0$. We consider two cases. (i) $C = 0$. Then, we have that $\ell^+(\mathbf{t}) = \ell^-(\mathbf{t}) = 0$, hence also $\ell(\mathbf{t}) = 0$, for any $\mathbf{t} \in N$ with $\nu^{\mathbf{T}}(\mathcal{T} \setminus N) = 0$. Thus, $\ell(\mathbf{T}(\mathbf{x})) = 0$ for any $\mathbf{x} \in \mathbf{T}^{-1}(N)$. Clearly,

$$\nu(\mathcal{X} \setminus \mathbf{T}^{-1}(N)) = \nu(\mathbf{T}^{-1}(\mathcal{T} \setminus N)) = \nu^{\mathbf{T}}(\mathcal{T} \setminus N) = 0,$$

so that $\mathrm{P}_{\boldsymbol{\theta}}[\mathcal{X} \setminus \mathbf{T}^{-1}(N)] = 0$ for any $\boldsymbol{\theta} \in \Theta$. In other words, $\ell(\mathbf{T}(\mathbf{X})) = 0$ $\mathrm{P}_{\boldsymbol{\theta}}$-a.s. for any $\boldsymbol{\theta} \in \Theta$. (ii) $C > 0$. Then, (4.5) implies that

$$f^{\pm}(\mathbf{t}) := \frac{1}{C} \ell^{\pm}(\mathbf{t}) \exp(\boldsymbol{\theta}_0'\mathbf{t})$$

are densities with respect to $\nu^{\mathbf{T}}$, whereas (4.4), which rewrites

$$\int_{\mathcal{T}} \exp((\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{t}) f^+(\mathbf{t}) d\nu^{\mathbf{T}}(\mathbf{t}) = \int_{\mathcal{T}} \exp((\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{t}) f^-(\mathbf{t}) d\nu^{\mathbf{T}}(\mathbf{t}) \quad \text{for any } \boldsymbol{\theta} \in \Theta,$$

implies that these densities have moment-generating functions that coincide in a neighborhood of the origin (recall that $\boldsymbol{\theta}_0$ is an interior point to $\boldsymbol{\Theta}$). Thus, $\ell^+(\mathbf{t}) = \ell^-(\mathbf{t})$ $\nu^{\mathbf{T}}$-almost everywhere, hence $\ell(\mathbf{t}) = 0$ $\nu^{\mathbf{T}}$-almost everywhere. The same reasoning as in case (i) then still yields that $\ell(\mathbf{T}(\mathbf{X})) = 0$ $\mathrm{P}_{\boldsymbol{\theta}}$-a.s. for any $\boldsymbol{\theta} \in \Theta$. □

Provided that a minimal sufficient statistic exists, $\mathbf{T}$, as a sufficient and complete statistic, is also minimal sufficient, so that one might think that Proposition 1 is a corollary of Proposition 2. It is not, though, since it was unclear before proving Proposition 1 that a minimal sufficient statistic does exist.

4.4 Further properties

(A) Combined with Propositions 1–2, Remark 5 in Section 4.1 is extremely important. Indeed, it implies that, if $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are identically and independently distributed random vectors from an exponential family with K-dimensional privileged statistic $\mathbf{T}(\mathbf{X})$, then the n-tuple $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is exponential, with K-dimensional privileged statistic $\sum_{i=1}^n \mathbf{T}(\mathbf{X}_i)$. *The dimension of the sufficient and complete statistic thus is K, irrespective of the sample size n.* Under appropriate regularity assumptions, it can be shown that exponential families are the only ones allowing for a sufficient statistic with fixed dimension K irrespective of the (possibly arbitrarily large) sample of size n—a result associated with the names of Darmois (1935), Koopman (1936) and Pitman (1937).

(B) In the rest of this chapter, we assume that the dominating measure was chosen to absorb the factor $h(\mathbf{x})$; see Remark 1. Assume then that $\phi : (\mathcal{X}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ is such that

$$\int \phi(\mathbf{x}) \exp\left(\sum_{k=1}^K \theta_k T_k(\mathbf{x})\right) \mathrm{d}\mu(\mathbf{x})$$

exists and is finite for all $\boldsymbol{\theta} \in \Theta$. Then,

(i) the mapping

$$\boldsymbol{\theta} \mapsto \int \phi(\mathbf{x}) \exp\left(\sum_{k=1}^K \theta_k T_k(\mathbf{x})\right) \mathrm{d}\mu(\mathbf{x}) \tag{4.6}$$

is *analytical* at all $\boldsymbol{\theta}$ in the interior of Θ_0;

(ii) its derivatives (of all orders) with respect to $\boldsymbol{\theta}$ can be computed by differentiating (4.6) under the integral sign.

(C) *Moments of the privileged statistic.* Consider the exponential family with densities

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = C(\boldsymbol{\theta}) \exp\left(\sum_{k=1}^{K} \theta_k T_k(\mathbf{x}) \right)$$

and let $\psi(\boldsymbol{\theta}) := -\log C(\boldsymbol{\theta})$.

Proposition 3. *For all $\boldsymbol{\theta}$ in the interior of Θ_0, all moments of $\mathbf{T}$ at $\mathrm{P}_{\boldsymbol{\theta}}$ exist and are finite. In particular, denoting as $\nabla_{\boldsymbol{\theta}}$ the gradient operator with respect to $\boldsymbol{\theta}$,*

$$\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{T}] = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) \quad \textit{and} \quad \mathrm{Var}_{\boldsymbol{\theta}}[\mathbf{T}] = \left(\partial^2_{\theta_i,\theta_j}\psi(\boldsymbol{\theta})\right) = \mathbf{I}(\boldsymbol{\theta}),$$

where

$$\mathbf{I}(\boldsymbol{\theta}) := \mathrm{E}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\mathbf{X}) \left(\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\mathbf{X})\right)']$$

is the Fisher information matrix for $\boldsymbol{\theta}$.

Proof. It follows from point (B) above that the moment-generating function of $\mathbf{T}$ is analytical; hence, all moments of $\mathbf{T}$ exist and are finite. For the remaining points in the proposition, we provide a direct derivation. Since $\int_{\mathcal{X}} f_{\boldsymbol{\theta}}(\mathbf{x})\mathrm{d}\mu(\mathbf{x}) = 1$, we have

$$\int_{\mathcal{X}} \exp\left(\sum_{k=1}^{K} \theta_k T_k(\mathbf{x}) \right) \mathrm{d}\mu(\mathbf{x}) = \frac{1}{C(\boldsymbol{\theta})} = \exp(\psi(\boldsymbol{\theta})).$$

In view of point (B) above, derivatives with respect to θ_i can be taken on both sides of that identity, under the integral sign, yielding, for all i,

$$\int_{\mathcal{X}} T_i(\mathbf{x}) \exp\left(\sum_{k=1}^{K} \theta_k T_k(\mathbf{x}) \right) \mathrm{d}\mu(\mathbf{x}) = \left(\partial_{\theta_i}\psi(\boldsymbol{\theta}) \exp(\psi(\boldsymbol{\theta}))\right). \tag{4.7}$$

Multiplying both sides with $C(\boldsymbol{\theta}) = \exp(-\psi(\boldsymbol{\theta}))$, we get

$$\mathrm{E}_{\boldsymbol{\theta}}[T_i] = \partial_{\theta_i}\psi(\boldsymbol{\theta}) \quad i = 1, \ldots, K.$$

Differentiating both sides of (4.7) with respect to θ_j now yields

$$\int_{\mathcal{X}} T_i(\mathbf{x})T_j(\mathbf{x}) \exp\left(\sum_{k=1}^{K} \theta_k T_k(\mathbf{x})\right) \mathrm{d}\mu(\mathbf{x}) = \left\{\left(\partial^2_{\theta_i,\theta_j}\psi(\boldsymbol{\theta})\right) + (\partial_{\boldsymbol{\theta}_i}\psi(\boldsymbol{\theta}))(\partial_{\boldsymbol{\theta}_j}\psi(\boldsymbol{\theta}))\right\} \exp(\psi(\boldsymbol{\theta})).$$

Then, multiplying both sides by $C(\boldsymbol{\theta}) = \exp(-\psi(\boldsymbol{\theta}))$,

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}}[T_iT_j] &= \partial^2_{\theta_i,\theta_j}\psi(\boldsymbol{\theta}) + (\partial_{\boldsymbol{\theta}_i}\psi(\boldsymbol{\theta}))(\partial_{\boldsymbol{\theta}_j}\psi(\boldsymbol{\theta})) \\ &= \partial^2_{\theta_i,\theta_j}\psi(\boldsymbol{\theta}) + \mathrm{E}_{\boldsymbol{\theta}}[T_i]\mathrm{E}_{\boldsymbol{\theta}}[T_j], \end{aligned}$$

which yields

$$\mathrm{Cov}_{\boldsymbol{\theta}}[T_i, T_j] = \partial^2_{\theta_i,\theta_j}\psi(\boldsymbol{\theta}).$$

Finally, since direct differentiation yields $\nabla_{\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\mathbf{x}) = -\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) + \mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{x}) - \mathrm{E}_{\boldsymbol{\theta}}[\mathbf{T}]$, we see that the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ exists and is finite, and that it coincides with $\mathrm{Var}_{\boldsymbol{\theta}}[\mathbf{T}]$, as was to be shown. □

It follows from Proposition 3 that $\mathbf{T}$, with covariance matrix $\mathbf{I}(\boldsymbol{\theta})$, moreover is an *efficient* estimator[3] of $\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{T}] = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})$, which (up to a linear transformation) is actually the only efficiently estimable function of $\boldsymbol{\theta}$.

[3] Recall that an estimator $\mathbf{S}$ is efficient for $\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{S}] = \boldsymbol{\Upsilon}(\boldsymbol{\theta}) = (\Upsilon_1(\boldsymbol{\theta}), \ldots, \Upsilon_K(\boldsymbol{\theta}))'$ if its covariance matrix reaches the Cramér-Rao bound $(\partial\Upsilon_i/\partial\theta_j)\,(\mathbf{I}(\boldsymbol{\theta}))^{-1}\,(\partial\Upsilon_i/\partial\theta_j)'$; for $\boldsymbol{\Upsilon}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})$, indeed, $(\partial\Upsilon_i/\partial\theta_j) = (\partial^2_{\theta_i,\theta_j}\psi(\boldsymbol{\theta})) = \mathbf{I}(\boldsymbol{\theta})$.