Exponential families

The family of distributions with range not depending on the parameter and with sufficient statistics that have dimension independent of sample size turns out to be quite rich. It is called the **exponential family** of distributions. These have density

$$f(\mathbf{x};\boldsymbol{\theta}) = \exp(\mathbf{c}(\boldsymbol{\theta})^T\mathbf{T}(\mathbf{x}) - B(\boldsymbol{\theta}))h(\mathbf{x}), \;\; \mathbf{x}\in E\subset\mathbf{R}^n$$

where the set E does not depend on $\boldsymbol{\theta}$. The sufficient statistic $\mathbf{T}(\mathbf{x})$, which is determined up to a multiplicative constant, is called the **natural sufficient statistic**. We say that an exponential family is **minimal** if the functions $\mathbf{c}(\boldsymbol{\theta})$ and the statistics $\mathbf{T}(\mathbf{x})$ each are linearly independent. We can always achieve this by reparametrization.

Example: (1) $\Gamma(\theta,1)$: For the gamma distribution with shape parameter θ and scale parameter 1 we have $T(\mathbf{x})=\sum\log x_i$, E=$\{x_i>0, i=1,\ldots,n\}$, $c(\theta)=\theta-1$, $B(\theta)=-n\log\Gamma(\theta)$ and $h(\mathbf{x})=\exp(-\sum x_i)$. □

(2) The inverse normal: This distribution is given by the density

$$f(\mu,\lambda)=\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}} x^{-3/2}\exp(-\lambda(x-\mu)^2/2\mu^2x), \;\; x>0.$$

Here μ is the mean, and λ is a precision parameter. It sometimes helps to reparametrize using $\alpha=\lambda/\mu^2$, yielding

$$f(\alpha,\lambda)=\exp((\alpha\lambda)^{\frac{1}{2}}-\tfrac{1}{2}\log\lambda-\alpha x/2-\lambda/2x)$$

so that $(\sum x_i, \sum x_i^{-1})$ is sufficient for the natural parameter $\alpha/2$, $\lambda/2$. □

For a minimal family, the sufficient statistic $\mathbf{T}$ is also minimal sufficient. For a proof, see Lehmann: *Theory of Point Estimation*, Example 5.9, pp. 43-44.

If we parametrize the family using $\boldsymbol{\eta}=c(\boldsymbol{\theta})$, this is called the **natural parametrization** (or the canonical parametrization). We then write

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

where

$$A(\boldsymbol{\eta})=\log\int_E\exp(\boldsymbol{\eta}^T\mathbf{T}(\mathbf{x}))h(\mathbf{x})d\mathbf{x}$$

The **natural parameter space** is $\mathbf{H}=\{\boldsymbol{\eta}: A(\boldsymbol{\eta})<\infty\}$.

Theorem 1: $\mathbf{H}$ is a convex set.

Proof: Let $0<\alpha<1$ and take $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_1$ in $\mathbf{H}$. Write

$$A(\alpha\boldsymbol{\eta}+(1-\alpha)\boldsymbol{\eta}_1)=\log\int_E(\exp(\boldsymbol{\eta}^T\mathbf{T}(\mathbf{x}))h((\mathbf{x})))^{\alpha}(\exp(\boldsymbol{\eta}_1^T\mathbf{T}(\mathbf{x}))h((\mathbf{x})))^{1-\alpha}d\mathbf{x}.$$

But $u^{\alpha}v^{1-\alpha}\le\alpha u+(1-\alpha)v$ (take logarithms of both sides and use the fact that the logarithm is a concave function), whence

$$A(\alpha\boldsymbol{\eta}+(1-\alpha)\boldsymbol{\eta}_1)\le\alpha A(\boldsymbol{\eta})+(1-\alpha)A(\boldsymbol{\eta}_1)<\infty.$$

Example: In the case of $\Gamma(\theta, 1)$ we define $h(\mathbf{x}) = \exp(-\sum x_i - \sum \log x_{i)}$, and find that θ is itself the natural parameter (obviously, any linear function of θ is also natural). The natural parameter space is then $\mathbf{R}_+$.
□

Theorem 2: If $d = \dim(\mathbf{H}) = 1$ we have whenever $\eta \in \mathrm{int}(\mathbf{H})$ that

$$\mathbf{E}_\eta T(\mathbf{X}) = A'(\eta)$$

and

$$\mathbf{Var}_\eta T(\mathbf{X}) = A''(\eta).$$

Proof: First compute

$$\mathbf{E}_\eta T(\mathbf{X}) = \int_E T(\mathbf{x}) f(\mathbf{x}; \eta) d\mathbf{x}$$

$$= \int_E T(\mathbf{x}) \exp(\eta T(\mathbf{x}) - A(\eta)) h(\mathbf{x}) d\mathbf{x}$$

and

$$A'(\eta) = \frac{d}{d\eta} \log \int_E \exp(\eta T) h d\mathbf{x} = \frac{1}{\int_E \exp(\eta T) h d\mathbf{x}} \int_E T \exp(\eta T) h d\mathbf{x}$$

$$= \int_E T \exp(\eta T - A) h d\mathbf{x} = \mathbf{E}_\eta T$$

where the differentiation under the integral sign needs to be be justified. To do that, define $\psi(\eta) = \exp(A(\eta))$. Note that

$$\frac{\psi(\eta + k) - \psi(\eta)}{k} = \frac{1}{k} \left(\int_E \exp((\eta + k) T) h d\mathbf{x} - \int_E \exp(\eta T) h d\mathbf{x} \right)$$

$$= \int_E \frac{\exp(kT) - 1}{k} \exp(\eta T) h d\mathbf{x}$$

Now, if $|k| < \delta$

$$\left| \frac{\exp(kT) - 1}{k} \right| = \left| \sum_{i=1}^{\infty} \frac{T^i k^{i-1}}{i!} \right| \le \sum_{i=1}^{\infty} \frac{|T|^i |k|^{i-1}}{i!} < \frac{\exp(|T|\delta) - 1}{\delta} < \frac{\exp(\delta T) + \exp(-\delta T)}{\delta}$$

and thus

$$\left| \frac{\psi(\eta + k) - \psi(\eta)}{k} \right| < \frac{1}{\delta} (\psi(\eta + \delta) + \psi(\eta - \delta)) < \infty$$

since $\eta \in \mathrm{int}(\mathbf{H})$. Hence the dominated convergence theorem applies, and

$$\lim_{k \to 0} \frac{\psi(\eta + k) - \psi(\eta)}{k} = \int_E \lim_{k \to 0} \frac{\exp(kT) - 1}{k} \exp(\eta T) h d\mathbf{x} = \int_E T \exp(\eta T) h d\mathbf{x}.$$

This justifies the interchange of integration and differentiation (and the argument can be repeated for higher derivatives).

To compute the variance, first note that

$$\frac{d^2}{ds^2}\log\psi(s) = \frac{\psi''(s)}{\psi(s)} - \left(\frac{\psi'(s)}{\psi(s)}\right)^2.$$

By differentiating under the integral sign $\psi''(\eta) = \int_E T^2 \exp(\eta T) h d\mathbf{x}$, so $\psi''/\psi = \mathbf{E}_\eta T^2$, whence

$$A'' = \frac{d^2}{d\eta^2}\log\psi(\eta) = \mathbf{E}T^2 - \mathbf{E}^2 T = \mathbf{Var}T.$$

□

Example: (1) $\Gamma(\theta, 1)$: For $n = 1$ we have $A(\theta) = \log\Gamma(\theta)$, so

$$\mathbf{E}_\theta T = \mathbf{E}_\theta \log X = \frac{\Gamma'(\theta)}{\Gamma(\theta)} = \Psi(\theta)$$

where $\Psi(x)$ is the digamma function.

(2) The Rayleigh density: let

$$f(x;\theta) = \frac{x}{\theta^2}\exp(-x^2/2\theta^2)$$

where x and θ are positive real numbers. This is the density of the length of a bivariate normal vector, where the two components are independent standard normal. Writing this in exponential family form for a sample of size n we get

$$f(\mathbf{x};\theta) = \exp(-\frac{1}{2\theta^2}\sum x_i^2 - n\log\theta^2 + \sum\log x_i).$$

Thus $\eta = c(\theta) = -1/2\theta^2$, or $\theta^2 = -1/2\eta$. Also $B(\theta) = n\log\theta^2$, so $A(\eta) = -n\log(-2\eta)$. By Theorem 1 we get

$$\mathbf{E}\sum X_i^2 = -\frac{n}{\eta} = 2n\theta^2$$

and

$$\mathbf{Var}\sum X_i^2 = \frac{n}{\eta^2} = 4n\theta^4.$$

To compute these moments directly from the density one must compute, e.g,

$$\int_0^\infty \frac{x^3}{\theta^2}\exp(-x^2/2\theta^2)dx,$$

an exercise in partial integration. □

Corollary 1: For $d > 1$ we have

$$\mathbf{E}_{\boldsymbol{\eta}} T_i(\mathbf{X}) = \frac{\partial}{\partial\eta_i}A(\boldsymbol{\eta})$$

and

$$\mathbf{Cov}_{\boldsymbol{\eta}}(T_i(\mathbf{X}), T_j(\mathbf{X})) = \frac{\partial^2}{\partial\eta_i\partial\eta_j}A(\boldsymbol{\eta})$$

Corollary 2: The moment generating function of $\mathbf{T}$ is

$$\mathbf{E}_{\eta} \exp(\sum t_i T_i) = \psi(\boldsymbol{\eta}+\mathbf{t})/\psi(\boldsymbol{\eta}),$$

where $\psi(\boldsymbol{\eta}) = \exp(A(\boldsymbol{\eta}))$.

Corollary 3: The cumulant generating function of T is

$$\kappa(\mathbf{t}) = \log \mathbf{E}_{\eta} \exp \sum t_i T_i = A(\boldsymbol{\eta}+\mathbf{t}) - A(\boldsymbol{\eta}).$$

Thus $A^{(k)}(\boldsymbol{\eta})$ is the k'th cumulant of T.

Example: (1) $\Gamma(\alpha, \beta)$: Here

$$f(\mathbf{x}, \boldsymbol{\eta}) = \exp(\alpha \sum \log x_i - \beta \sum x_i - n\alpha \log \beta - n \log \Gamma(\alpha) - \sum \log x_i)$$

The natural parameter is $\boldsymbol{\eta} = (\alpha, \beta)^T$, with $A(\eta) = \alpha \log \beta + \log \Gamma(\alpha)$, and by Corollary 2 the joint mgf for $(\log X, X)$ is

$$\exp(\log(\Gamma(\alpha+t_1) + (\alpha+t_1)\log(\beta+t_2)) - \log\Gamma(\alpha) - \alpha \log \beta) = \frac{\Gamma(\alpha+t_1)}{\Gamma(\alpha)} (1+t_2/\beta)^{\alpha} (\beta+t_2)^{t_1}$$

Setting $t_1 = 0$ we get the mgf for X, and setting $t_2 = 0$ that for $\log X$. □

(2) Inverse normal: By Corollary 3, the cgf for the inverse normal distribution is $\kappa(\alpha, \lambda) = \frac{1}{2}\log\lambda + (\alpha\lambda)^{\frac{1}{2}}$. □

The exponential family is closed under sampling, in the sense that if we take a random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from the distribution (1), the joint density is

$$\prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta}) = \exp(\sum c(\boldsymbol{\theta})^T \mathbf{T}(\mathbf{x}_i) - nB(\boldsymbol{\theta})) \prod h(\mathbf{x}_i) = \exp(c(\boldsymbol{\theta})^T (\sum \mathbf{T}(\mathbf{x}_i)) - nB(\boldsymbol{\theta})) h_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

so the natural sufficient statistic is $\sum \mathbf{T}(\mathbf{x}_i)$. Notice that this always has the same dimension, regardless of sample size. Furthermore, the natural sufficient statistic itself has an exponential family distribution. More precisely, we have the following result:

Theorem 3: Let X be distributed according to the exponential family

$$f(\mathbf{x}; (\boldsymbol{\alpha}, \boldsymbol{\beta})^T) = \exp(\sum_{i=1}^{r} \alpha_i U_i(\mathbf{x}) + \sum_{i=1}^{s} \beta_i T_i(\mathbf{x}) - A(\alpha, \beta)) h(\mathbf{x}).$$

Then the distribution of $T_1, \ldots, T_s$ is an exponential family of the form

$$f_{\mathbf{T}}(\mathbf{t}; (\boldsymbol{\alpha}, \boldsymbol{\beta})^T) = \exp(\sum_{i=1}^{s} \beta_i t_i - A(\alpha, \beta)) g(\mathbf{t}; \boldsymbol{\alpha})$$

and the conditional distribution of $U_1, \ldots, U_r$ given $\mathbf{T} = \mathbf{t}$ is an exponential family of the form

$$f_{\mathbf{U}|\mathbf{T}}(\mathbf{u}|\mathbf{t}; (\boldsymbol{\alpha}, \boldsymbol{\beta})^T) = \exp(\sum_{i=1}^{r} \alpha_i u_i - A_{\mathbf{t}}(\alpha)) k(\mathbf{t}, \mathbf{u}).$$

Proof: We prove the result in the discrete case. A general proof is in Lehmann: *Testing Statistical Hypotheses*, Wiley, 1959, p.52. Compute

$$\mathbf{P}(\mathbf{T}(\mathbf{X}) = \mathbf{t}) = \sum_{\mathbf{x}:\mathbf{T}(\mathbf{x})=\mathbf{t}} f(\mathbf{x}; (\boldsymbol{\alpha}, \boldsymbol{\beta})^T))$$

$$= \sum_{\mathbf{x}:\mathbf{T}(\mathbf{x})=\mathbf{t}} \exp(\sum_{i=1}^{r} \alpha_i U_i(\mathbf{x}) + \sum_{i=1}^{s} \beta_i T_i(\mathbf{x}) - A(\alpha, \beta)) h(\mathbf{x})$$

$$= \exp(\sum_{i=1}^{s} \beta_i t_i - A(\alpha, \beta)) \sum_{\mathbf{x}:\mathbf{T}(\mathbf{x})=\mathbf{t}} \exp(\sum_{i=1}^{r} \beta_i U_i(\mathbf{x}))h(\mathbf{x})$$

This proves the first part. For the second part, write

$$\mathbf{P}(\mathbf{U} = \mathbf{u} | \mathbf{T} = \mathbf{t}) = \mathbf{P}(\mathbf{U} = \mathbf{u}, \mathbf{T} = \mathbf{t}) / \mathbf{P}(\mathbf{T} = \mathbf{t})$$

$$= \frac{\exp(\sum_{i=1}^{r} \alpha_i u_i + \sum_{i=1}^{s} \beta_i t_i - A(\alpha, \beta))k(\mathbf{u}, \mathbf{t})}{\exp(\sum_{i=1}^{s} \beta_i t_i - A(\alpha, \beta))g(\mathbf{t}; \alpha)} = \exp(\sum \alpha_i u_i - \log g(\mathbf{t}; \alpha))k(\mathbf{u}, \mathbf{t})$$

where $k(\mathbf{u}, \mathbf{t}) = \sum_{\mathbf{x}:\mathbf{U}(\mathbf{x})=\mathbf{u},\mathbf{T}(\mathbf{x})=\mathbf{t}} h(\mathbf{x})$. □

The likelihood equation for an exponential family is simple.

Theorem 4: Let $\mathbf{C} = \mathrm{int}(\mathbf{c}(\boldsymbol{\theta}): \boldsymbol{\theta} \in \Theta)$, and let $\mathbf{X}$ be distributed according to a minimal exponential family. If the equation

$$\mathbf{E}_{\boldsymbol{\theta}} \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})$$

has a solution $\hat{\boldsymbol{\theta}}(\mathbf{x})$ with $\mathbf{c}(\hat{\boldsymbol{\theta}}(\mathbf{x})) \in \mathbf{C}$, then $\hat{\boldsymbol{\theta}}$ is the unique mle of $\boldsymbol{\theta}$.

Proof: Look first at the canonical parametrization, so that the log likelihood is

$$\sum \eta_i T_i(\mathbf{x}) - A(\boldsymbol{\eta}) - \log h(\mathbf{x})$$

and the likelihood equations are

$$T_i(\mathbf{x}) = \frac{\partial}{\partial \eta_i} A(\boldsymbol{\eta}) = \mathbf{E}_{\boldsymbol{\eta}} T_i(\mathbf{X})$$

using Corollary 1 to Theorem 1. In addition, minus the second derivative of the log likelihood is just $[\partial A(\boldsymbol{\eta}) / \partial \eta_i \partial \eta_j]$, or the covariance matrix of $\mathbf{T}$. Since the parametrization is minimal, the covariance matrix is positive definite.

Now note that the values of $l_n(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$ are a subset of the values of $l_n(\boldsymbol{\eta}), \boldsymbol{\eta} \in \mathbf{H}$. Hence if $\mathbf{c}(\hat{\theta}) \in \mathbf{C}$ then this corresponds to the unique maximizing $\hat{\boldsymbol{\eta}}$. □

Example: $\Gamma(\alpha, \beta)$: The likelihood equations are

$$n \frac{\alpha}{\beta} = \sum x_i \quad n(\log \beta - \Psi(\alpha)) = \sum \log x_i$$

□