Exponential Families

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

Surprisingly many of the distributions we use in statistics for random variables X taking value in some space $\mathcal{X}$ (often $\mathbb{R}$ or $\mathbb{N}_0$ but sometimes $\mathbb{R}^n$, $\mathbb{Z}$, or some other space), indexed by a parameter θ from some parameter set Θ, can be written in **exponential family** form, with pdf or pmf

$$f(x \mid \theta) = \exp\left[\eta(\theta)t(x) - B(\theta)\right]\; h(x)$$

for some **statistic** $t : \mathcal{X} \to \mathbb{R}$, **natural parameter** $\eta : \Theta \to \mathbb{R}$, and functions[1] $B : \Theta \to \mathbb{R}$ and $h : \mathcal{X} \to \mathbb{R}_+$. The likelihood function for a random sample of size n from the exponential family is

$$f_n(\mathbf{x} \mid \theta) = \exp\left[\eta(\theta)\sum_{j=1}^{n} t(x_j) - nB(\theta)\right] \prod h(x_i),$$

which is actually of the same form with the same natural parameter $\eta(\cdot)$, but now with statistic $T_n(\mathbf{x}) = \sum t(x_j)$ and functions $B_n(\theta) = nB(\theta)$ and $h_n(\mathbf{x}) = \Pi h(x_j)$.

Examples

For example, the pmf for the binomial distribution $\mathsf{Bi}(m, p)$ can be written as

$$\binom{m}{x} p^x (1-p)^{m-x} = \exp\left[\left(\log\frac{p}{1-p}\right)x + m\log(1-p)\right]\binom{m}{x}$$

[1] For students acquainted with measure-theoretic probability: more generally, we can replace the function $h(x)$ with an arbitrary reference measure $h(dx)$ on $\mathcal{X}$, leading to the distribution measure $f(dx \mid \theta)$ for X. This lets us treat discrete and continuous distributions together.

of Exponential Family form with natural parameter $\eta(p) = \log \frac{p}{1-p}$ and natural sufficient statistic $t(x) = x$, and the Poisson

$$\frac{\theta^x}{x!} e^{-\theta} = \exp\left[(\log \theta) x - \theta\right] \, \frac{1}{x!}$$

with $\eta = \log \theta$ and again $t(x) = x$. The Beta distribution $\mathsf{Be}(\alpha, \beta)$ with either *one* of its two parameters unknown can be written in EF form too:

$$\begin{aligned} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} &= \exp\left[\alpha \log x - \left(\log \frac{\Gamma(\alpha)}{\Gamma(\alpha+\beta)}\right)\right] \frac{(1-x)^{\beta}}{x(1-x)\Gamma(\beta)} \\ &= \exp\left[\beta \log(1-x) - \left(\log \frac{\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right)\right] \frac{x^{\alpha}}{x(1-x)\Gamma(\alpha)} \end{aligned}$$

with $t(x) = \log x$ or $\log(1-x)$ when $\eta = \alpha$ or $\eta = \beta$ is unknown, respectively. With *both* parameters unknown the beta distribution can be written as a *bivariate* Exponential Family with parameter $\theta = [\alpha, \beta]' \in \mathbb{R}_+^2$:

$$f(x \mid \theta) = \exp\left[\eta(\theta) \cdot t(x) - B(\theta)\right] h(x) \tag{1}$$

with *vector* parameter $\eta = [\alpha, \beta]'$ and statistic $t(x) = [\log x, \log(1-x)]'$ and scalar (one-dimensional) functions $B(\theta) = \log \Gamma(\alpha) + \log \Gamma(\beta) - \log \Gamma(\alpha+\beta)$ and $h(x) = 1/x(1-x)$. Since this comes up often, we'll let η and T be q-dimensional below; usually in this course $q = 1$ or 2.

Natural Exponential Families

It is often convenient to reparameterize exponential families to the *natural parameter* $\eta = \eta(\theta) \in \mathbb{R}^q$, leading (with $A(\eta(\theta)) \equiv B(\theta)$) to

$$f(x \mid \eta) = e^{\eta \cdot t(x) - A(\eta)} h(x) \tag{2}$$

Since any pdf integrates to unity we have

$$e^{A(\eta)} = \int_{\mathcal{X}} e^{\eta \cdot t(x)} h(x) \, dx$$

and hence can calculate the moment generating function (MGF) for the **natural sufficient statistic** $t(x) = \{t_1(x), \cdots, t_q(x)\}$ as

$$
\begin{aligned}
M_t(s) &= \mathsf{E}\left[e^{s\cdot t(X)}\right] \\
&= \int_{\mathcal{X}} e^{s\cdot t(x)}\, e^{\eta\cdot t(x) - A(\eta)} h(x)\, dx \\
&= e^{-A(\eta)} \int_{\mathcal{X}} e^{(\eta+s)\cdot t(x)} h(x)\, dx \\
&= e^{A(\eta+s)-A(\eta)},
\end{aligned}
$$

so $\log M_t(s) = A(\eta + s) - A(\eta)$ and we can find moments for the natural sufficient statistic by

$$
\begin{aligned}
\mathsf{E}[t] &= \nabla \log M_t(0) &= \nabla A(\eta) \\
\mathsf{V}[t] &= \nabla^2 \log M_t(0) &= \nabla^2 A(\eta)
\end{aligned}
$$

provided that η is an interior point of the *natural parameter space*

$$
\mathcal{E} \equiv \{\eta \in \mathbb{R}^q : 0 < \int_{\mathcal{X}} e^{\eta\cdot t(x)} h(x)\, dx < \infty\}
$$

and that $A(\cdot)$ is twice-differentiable near η. For samples of size $n \in \mathbb{N}$ the sufficient statistic

$$
T_n(\mathbf{x}) = \sum t(x_j)
$$

is a sum of independent random variables, so by the Central Limit Theorem we have approximately

$$
\sim \mathsf{No}\Big(n\nabla A(\eta),\ n\nabla^2 A(\eta)\Big).
$$

Note that $\nabla^2 A(\eta) = -\nabla^2 \log f(\mathbf{x} \mid \theta)$ is both the observed and Fisher (expected) information (matrix) $I_n(\theta)$ for natural exponential families, and that the score statistic is $Z := \nabla \log f(\mathbf{x} \mid \theta) = \big[T_n(\mathbf{x}) - n\nabla A(\eta)\big]$.

Conjugate Priors

Fix a nonnegative function[2] $\pi_\star(\theta)$ on Θ and let $\mathcal{E}_\star \subseteq \mathbb{R}^{q+1}$ be the collection of hyper-parameter pairs (α, β) with $\alpha \in \mathbb{R}^q$, $\beta \in \mathbb{R}$ for which

$$
0 < c_{\alpha,\beta} := \int_{\Theta} e^{\eta(\theta)\cdot\alpha - \beta B(\theta)}\, \pi_\star(\theta)\, d\theta < \infty.
$$

[2] Again, an arbitrary positive reference measure $\pi_\star(d\theta)$ on Θ can replace the function $\pi_\star(\theta)$ here, leading to prior and posterior distributions that may not have Lebesgue densities, or that may be supported on a lower-dimensional subset of Θ.

We can define a $(q+1)$-dimensional parametric family of prior densities for $(\alpha, \beta) \in \mathcal{E}_\star$ by

$$\pi(\theta \mid \alpha, \beta) := c_{\alpha,\beta}^{-1} e^{\eta(\theta)\cdot\alpha - \beta B(\theta)}\, \pi_\star(\theta).$$

With this prior and with data $\{X_i\} \overset{\text{iid}}{\sim} f(x \mid \theta)$ from the exponential family, the posterior pdf is

$$\begin{aligned} \pi(\theta \mid \mathbf{x})) &\propto e^{\eta(\theta)\cdot\alpha - \beta B(\theta)} e^{\eta(\theta)\cdot T_n(\mathbf{x}) - nB(\theta)}\, \pi_\star(\theta) \\ &\propto \pi(\theta \mid \alpha^* = \alpha + T_n(\mathbf{x}),\ \beta^* = \beta + n). \end{aligned}$$

provided that $(\alpha^*, \beta^*) \in \mathcal{E}_\star$. This is within the same conjugate family but now with "updated" parameters $\alpha^* = \alpha + T_n$ and $\beta^* = \beta + n$. For example, in the binomial example above with constant $\pi_\star(p) \equiv 1$ on the unit interval this conjugate prior family has density function

$$\pi(p \mid \alpha, \beta) \propto \exp\left\{\alpha \log \frac{p}{1-p} - \beta \log(1-p)\right\} = p^\alpha (1-p)^{-(\alpha+\beta)},$$

the Beta family, with $\mathcal{E}_\star = \{\alpha, \beta :\ \alpha > -1,\ (\alpha+\beta) < 1\}$ while for the Poisson example it is

$$\pi(\theta \mid \alpha, \beta) \propto \exp\{\alpha \log \theta - \beta\theta\} = \theta^\alpha e^{-\beta\theta} \mathbf{1}_{\{\theta > 0\}}$$

for $\alpha > -1$ and $\beta > 0$, the Gamma family. Conjugate families for every exponential family are available in the same way.

Note not *every* distribution we consider is from an exponential family. From (2), for example, it is clear set of points where the pdf or pmf is nonzero, the possible values a random variable X can take, is just

$$\{x \in \mathcal{X} :\ f(x \mid \theta) > 0\} = \{x \in \mathcal{X} :\ h(x) > 0\},$$

which does *not* depend on the parameter θ; thus any family of distributions where the "support" depends on the parameter (uniform distributions are important examples, or location-scale families made from Gamma or Pareto distributions) can't be from an exponential family.

The table starting on page 6 show several familiar (and some less familiar ones, like the Inverse Gaussian $\mathsf{IG}(\mu, \lambda)$ and Pareto $\mathsf{Pa}(\alpha, \beta)$) distributions in exponential family form. Some of the formulas involve the log gamma function $\gamma(z) := \log\Gamma(z)$ and its first and second derivatives, the "digamma" $\psi(z) := (d/dz)\gamma(z)$ and "trigamma" $\psi'(z) := (d^2/dz^2)\gamma(z)$, which are built into `R`, Mathematica, Maple, the `gsl` library in C, and such, but aren't on pocket calculators or most spreadsheets. In each case $\nabla^2 A(\eta)$ is the Information matrix in the natural parameterization, $I(\theta)$ in the usual parameterization.

1 Exponential Family Examples

Be(α,β)

$$
\begin{aligned}
f(x) &= \tfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in (0,1) & T &= (\log x, \log 1-x)\\
B(\alpha,\beta) &= \gamma(\alpha)+\gamma(\beta)-\gamma(\alpha+\beta) & \eta &= (\alpha,\beta)\\
A(\eta) &= \gamma(\eta_1)+\gamma(\eta_2)-\gamma(\eta_1+\eta_2) & &\\
\nabla A(\eta) &= \begin{bmatrix}\psi(\eta_1)-\psi(\eta_1+\eta_2)\\ \psi(\eta_2)-\psi(\eta_1+\eta_2)\end{bmatrix} & \mathsf{E}T &= \begin{bmatrix}\psi(\alpha)-\psi(\alpha+\beta)\\ \psi(\beta)-\psi(\alpha+\beta)\end{bmatrix}\\
\nabla^2 A(\eta) &= \begin{pmatrix}\psi'(\eta_1)-c & -c\\ -c & \psi'(\eta_2)-c\end{pmatrix} & c &= \psi'(\eta_1+\eta_2)
\end{aligned}
$$

Bi(m,p)

$$
\begin{aligned}
f(x) &= \tbinom{m}{x} p^x q^{(m-x)}, \quad x = 0...m & T &= x\\
B(p) &= -m\log q & \eta &= \log(p/q)\\
A(\eta) &= m\log(1+e^\eta) & p &= e^\eta/(1+e^\eta)\\
\nabla A(\eta) &= \tfrac{me^\eta}{1+e^\eta} & \mathsf{E}T &= m\,p\\
\nabla^2 A(\eta) &= \tfrac{m\,e^\eta}{(1+e^\eta)^2} & I(p) &= m/pq
\end{aligned}
$$

Ex(λ)

$$
\begin{aligned}
f(x) &= \lambda e^{-\lambda x}, \quad x > 0 & T &= x\\
B(\lambda) &= -\log\lambda & \eta &= -\lambda\\
A(\eta) &= -\log(-\eta) & &\\
\nabla A(\eta) &= -1/\eta & \mathsf{E}T &= 1/\lambda\\
\nabla^2 A(\eta) &= \eta^{-2} & I(\lambda) &= 1/\lambda^2
\end{aligned}
$$

Ga(α,λ)

$$
\begin{aligned}
f(x) &= \tfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\, e^{-\lambda x}, \quad x > 0 & T &= (\log x, x)\\
B(\alpha,\beta) &= \gamma(\alpha)-\alpha\log\lambda & \eta &= (\alpha,-\lambda)\\
A(\eta) &= \gamma(\eta_1)-\eta_1\log(-\eta_2) & &\\
\nabla A(\eta) &= \begin{bmatrix}\psi(\eta_1)-\log(-\eta_2)\\ -\eta_1/\eta_2\end{bmatrix} & \mathsf{E}T &= \begin{bmatrix}\psi(\alpha)-\log\lambda\\ \alpha/\lambda\end{bmatrix}\\
\nabla^2 A(\eta) &= \begin{pmatrix}\psi'(\eta_1) & -1/\eta_2\\ -1/\eta_2 & \eta_1/{\eta_2}^2\end{pmatrix} & I(\alpha,\lambda) &= \begin{pmatrix}\psi'(\alpha) & -1/\lambda\\ -1/\lambda & \alpha/\lambda^2\end{pmatrix}
\end{aligned}
$$

Ge(p)

$$
\begin{aligned}
f(x) &= p\,q^x, \quad x = 0,1,2,... & T &= x\\
B(p) &= -\log p & \eta &= \log q\\
A(\eta) &= -\log(1-e^\eta) & p &= 1-e^\eta\\
\nabla A(\eta) &= \tfrac{e^\eta}{1-e^\eta} & \mathsf{E}T &= q/p\\
\nabla^2 A(\eta) &= \tfrac{e^\eta}{(1-e^\eta)^2} & I(p) &= 1/p^2 q
\end{aligned}
$$

Exponential Family Examples (cont'd)

$\mathsf{IG}(a,b)$

$$\begin{aligned} f(x) &= a e^{-(a-bx)^2/2x}/\sqrt{2\pi x^3}, \quad x>0 & T &= (1/x, x)^\intercal \\ B(a,b) &= -ab - \log a & \eta &= (-a^2/2, -b^2/2)^\intercal \\ A(\eta) &= -2\sqrt{\eta_1\,\eta_2} - \tfrac{1}{2}\log(-2\eta_1) & a &= \sqrt{-2\eta_1}, \quad b = \sqrt{-2\eta_2} \\ \nabla A(\eta) &= \begin{bmatrix} \sqrt{\eta_2/\eta_1} - 1/2\eta_1 \\ \sqrt{\eta_1/\eta_2} \end{bmatrix} & \mathsf{E}T &= \begin{bmatrix} b/a + 1/a^2 \\ a/b \end{bmatrix} \\ \nabla^2 A(\eta) &= \tfrac{1}{2}\begin{pmatrix} \sqrt{\frac{\eta_2}{\eta_1{}^3}} + \frac{1}{\eta_1{}^2} & \frac{-1}{\sqrt{\eta_1\eta_2}} \\ \frac{-1}{\sqrt{\eta_1\eta_2}} & \sqrt{\frac{\eta_1}{\eta_2{}^3}} \end{pmatrix} & I(a,b) &= \begin{pmatrix} b/a + 2/a^2 & -1 \\ -1 & a/b \end{pmatrix} \end{aligned}$$

$\mathsf{NB}(\alpha,p)$

$$\begin{aligned} f(x) &= \tbinom{-\alpha}{x} p^\alpha (-q)^x, \quad x = 0,1,2,\ldots & T &= x \\ B(p) &= -\alpha\log p & \eta &= \log q \\ A(\eta) &= -\alpha\log(1-e^\eta) & p &= 1-e^\eta \\ \nabla A(\eta) &= \tfrac{\alpha e^\eta}{1-e^\eta} & \mathsf{E}T &= \alpha q/p \\ \nabla^2 A(\eta) &= \tfrac{\alpha e^\eta}{(1-e^\eta)^2} & I(p) &= \alpha/p^2 q \end{aligned}$$

$\mathsf{No}(\mu,\sigma^2)$

$$\begin{aligned} f(x) &= e^{-(x-\mu)^2/2\sigma^2}/\sqrt{2\pi\sigma^2} & T &= (x, x^2)^\intercal \\ B(\mu,\sigma^2) &= \mu^2/2\sigma^2 + \tfrac{1}{2}\log\sigma^2 & \eta &= (\mu\sigma^{-2}, -\sigma^{-2}/2)^\intercal \\ A(\eta) &= -\eta_1{}^2/4\eta_2 - \tfrac{1}{2}\log(-2\eta_2) & (\mu,\sigma^2)^\intercal &= -(\eta_1, 1)^\intercal/2\eta_2 \\ \nabla A(\eta) &= \begin{bmatrix} -\eta_1/2\eta_2 \\ \eta_1{}^2/4\eta_2{}^2 - 1/2\eta_2 \end{bmatrix} & \mathsf{E}T &= \begin{bmatrix} \mu \\ \mu^2+\sigma^2 \end{bmatrix} \\ \nabla^2 A(\eta) &= \begin{pmatrix} -1/2\eta_2 & \eta_1/2\eta_2{}^2 \\ \eta_1/2\eta_2{}^2 & -\eta_1{}^2/2\eta_2{}^3 + 1/2\eta_2{}^2 \end{pmatrix} & I(a,b) &= \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & \sigma^{-4}/2 \end{pmatrix} \end{aligned}$$

$\mathsf{Po}(\lambda)$

$$\begin{aligned} f(x) &= \lambda^x e^{-\lambda}/x!, \quad x = 0,1,2,\ldots & T &= x \\ B(\lambda) &= \lambda & \eta &= \log\lambda \\ A(\eta) &= e^\eta & \lambda &= e^\eta \\ \nabla A(\eta) &= e^\eta & \mathsf{E}T &= \lambda \\ \nabla^2 A(\eta) &= e^\eta & I(\lambda) &= 1/\lambda \end{aligned}$$

$\mathsf{Pa}(\alpha,\beta)$

$$\begin{aligned} f(x) &= \beta\,\alpha^\beta/x^{\beta+1}, \quad x>\alpha & T &= \log x \\ B(\beta) &= -\log\beta - \beta\log\alpha & \eta &= -\beta \\ A(\eta) &= -\log(-\eta) + \eta\log\alpha & \beta &= -\eta \\ \nabla A(\eta) &= \log\alpha - 1/\eta & \mathsf{E}T &= \log\alpha + 1/\beta \\ \nabla^2 A(\eta) &= \eta^{-2} & I(\lambda) &= \beta^{-2} \end{aligned}$$