Statistical Approaches to Learning and Discovery

Week 2: Some Basics of Exponential Families and Bayesian Inference

January 22, 2003

The Exponential Family

Setup:

- Sample space $\mathcal{X}$ (σ-finite measure μ for $\mathcal{X}$)
- Statistics $t_k : \mathcal{X} \longrightarrow \mathbb{R}$, $k = 1, 2, \ldots, n$
- Default, or carrier density p_0 on $\mathcal{X}$, $p_0(x) \geq 0$

$(\mathcal{X}, \mu, t, p_0)$ determines an exponential family

$$p(x \mid \theta) = p_0(x)\, e^{\langle \theta, t(x) \rangle - \psi(\theta)}$$

The Exponential Family (cont.)

density $\quad p(x \mid \theta) = p_0(x)\ \exp\left(\langle \theta, t(x) \rangle - \psi(\theta)\right)$

cumulant function $\quad \psi(\theta) = \log \int_{\mathcal{X}} p_0(x)\, e^{\langle \theta, t(x) \rangle} d\mu(x) = \log Z(\theta)$

natural param. space $\quad \Theta = \left\{ \theta \in \mathbb{R}^n \mid \int_{\mathcal{X}} p_0(x)\, e^{\langle \theta, t(x) \rangle} \mu(dx) < \infty \right\}$

Example: Gaussian

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Completing square, find that density has parameterization

$$f(x \mid \theta) = \exp\left(\theta_1\, x + \theta_2\, x^2 - \Psi(\theta)\right)$$

Sufficient statistics $t_1(x) = x$ and $t_2(x) = x^2$.

In terms of standard param., $\theta_1 = \frac{\mu}{\sigma^2}$ and $\theta_2 = -\frac{1}{2\sigma^2}$

Dimension two. Domain (natural parameter space)

$$\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 \mid \theta_2 < 0\}$$

Calculus of the Exponential Family

Mean and variance obtained from derivatives of the cumulant function:

$$\begin{aligned}
\frac{\partial \psi(\theta)}{\partial \theta_k} &= \frac{\partial \log Z(\theta)}{\partial \theta_k} \\
&= \frac{1}{Z(\theta)} \int_{\mathcal{X}} p_0(x) e^{\langle \theta, t(x) \rangle} t_k(x)\, \mu(dx) \\
&= E_\theta[t_k] \\
\frac{\partial^2 \psi(\theta)}{\partial \theta_j \partial \theta_k} &= E_\theta[t_j\, t_k] - E_\theta[t_j]\, E_\theta[t_k]
\end{aligned}$$

Calculus of the Exponential Family (cont.)

Fisher information matrix $J(\theta)$, variance of the score:

$$\begin{aligned}
J_{jk}(\theta) &= E_\theta[\frac{\partial \log p(X \mid \theta)}{\partial \theta_j} \frac{\partial \log p(X \mid \theta)}{\partial \theta_k}] \\
&= E_\theta\left[(t_j - E_\theta[t_j])(t_k - E_\theta[t_k])\right] \\
\Rightarrow \quad J(\theta) &= \nabla^2 \psi(\theta)
\end{aligned}$$

$\nabla^2 \psi$ is positive-definite (assuming $\dim \Theta = n$)

They're Everywhere

Exponential models are everywhere – though often a reduction by sufficiency & reparameterization is required

Standard example: Bernoulli

$$p(x \mid \pi) = \pi^x (1-\pi)^{1-x}$$

Let $\theta_1 = \log \pi,\ \theta_2 = \log(1-\pi)$

Then $p(x \mid \theta) = e^{\theta_1 x + \theta_2 (1-x)}$

This not a natural (canonical) parameterization, because there is a constraint:

$$e^{\theta_1} + e^{\theta_2} = 1$$

Standard Parameterization of a Bernoulli

$$\begin{aligned}
\theta &= \log\left(\frac{\pi}{1-\pi}\right) \\
t(x) &= x \\
p_0(x) &= 1 \\
\psi(\theta) &= \log\left(1 + e^\theta\right) \\
\pi &= \frac{1}{1+e^{-\theta}} \quad \text{(logistic)}
\end{aligned}$$

Logistic Parameterization of a Multinomial

$$
\begin{aligned}
p(x_1, \ldots, x_{p+1}) &= \pi_1^{x_1} \cdots \pi_{p+1}^{x_{p+1}} \\
\theta_k &= \begin{cases} \log \frac{\pi_k}{\pi_{p+1}} & k \neq p+1 \\ 0 & k = p+1 \end{cases} \\
t_k(x) &= \begin{cases} x_k & k \neq p+1 \\ 0 & k = p+1 \end{cases} \\
\psi(\theta) &= \log \left(1 + \sum_{k=1}^{p} e^{\theta_k} \right)
\end{aligned}
$$

There are other possible parameterizations...

Expectation Parameterization

We've seen that

$$g(\theta) \stackrel{\text{def}}{=} \nabla\psi(\theta) = E_\theta[t] \stackrel{\text{def}}{=} \mu$$

The cumulant function is convex, and so we can invert this.

Many distributions are usually parameterized in terms of the expectation parameter μ.

We'll return to this when we talk about duality.

Maximum Likelihood Estimation

Again under some "regularity" conditions, the likelihood function is strictly concave for exponential families, and the MLE exists and is unique.

Several common exponential family models have closed form MLEs...examples?

However, calculating the MLE generally involves numerical methods. For large scale problems, the particular numerical methods chosen can be important.

Moment equations characterizing the MLE:

$$\frac{1}{N} \sum_{i=1}^{N} t_k(x_i) = \int_{\mathcal{X}} t_k(x)\, p(x \,|\, \theta)\, d\mu(x)$$

Hybrid Exponential Models

Can use the carrier density p_0 to form a kind of semi-parametric exponential family.

Take $\hat{p}_0(x)$ to be a kernel density estimate (for example). Then, form model

$$p(x \,|\, \theta) = \hat{p}_0(x)\, \exp\left(\langle \theta, t(x) \rangle - \psi(\theta)\right)$$

where θ is fit to maximize likelihood—e.g., to match moments

$$\int t(x) p(x \,|\, \theta)\, dx = \frac{1}{N} \sum_{i=1}^{N} t(x_i)$$

Hybrid Exponential Models (cont.)

Two views: exponential model with data-dependent p_0, or non-parametric density estimate "corrected" to match moments

B. Efron and R. Tibshirani, "Using specially designed exponential families for density estimation," *Annals of Statistics*, 24(6), pp. 2431–2461, 1996

Conjugacy (will return to this...)

Write our exponential model in the form (by reduction from sufficiency)

$$p(x \,|\, \theta) \;=\; e^{\langle \theta, x\rangle - \psi(\theta)}$$

A prior is *conjugate* if it is closed under sampling. Conjugate priors have the following form:

$$p(\theta \,|\, \alpha, \gamma) = c e^{\langle \theta, \gamma\rangle - \alpha\psi(\theta)}$$

for some $\alpha \in \mathbb{R}$ and $\gamma \in \mathbb{R}^d$.

Conjugacy (cont.)

More suggestive form:

$$p(\theta \,|\, n_0, x_0) \;=\; c e^{n_0\langle \theta, x_0\rangle - n_0\psi(\theta)}$$

Let $X_1, \ldots, X_n$ be a sample from $p(\cdot \,|\, \theta)$ under this prior. Then the posterior distribution is

$$p\left(\theta \;\middle|\; n_0 + n,\; \tfrac{n_0 x_0 + n\overline{X}}{n_0 + n}\right)$$

where $\overline{X}$ is the sample mean.

Conjugacy (cont.)

The conjugate prior is suitable for Bayesian estimation of the *expectation* parameters, due to the following property:

$$E[\nabla\psi] = \int_\Theta \nabla\psi(\theta)\, p(\theta \,|\, n_0, x_0)\, d\theta = x_0$$

The converse is also true: the posterior expectation $E[\nabla\psi \,|\, X_1]$ is linear in X_1 if and only if

$$p(\theta) = c e^{\langle \theta, \gamma\rangle - \alpha\psi(\theta)}$$

Persi Diaconis and Donald Ylvisaker, "Conjugate priors for exponential families," *Annals of Statistics* 7, 269-281, 1979.

Conjugacy (cont.)

Conjugacy is often motivated by computational considerations.

However, recent results give additional justification in terms of independence properties.

Conjugacy is typically not needed (or desired) when using MCMC methods, as we'll discuss.

D. Geiger, D. Heckerman, "A characterization of the Dirichlet distribution through global and local independence," *Annals of Statistics* 25, pp. 1344–1369, 1997.