Chapter 3: Exponential Families and Conjugate Priors

Contents

1 Introduction

Exponential families are a unifying generalization of many basic probabilistic models, and they possess many special properties. In fact, we have already encountered several exponential families—Bernoulli, Beta, Exponential, and Gamma—and there are many more. From the Bayesian perspective, a key feature of exponential families is that often the posterior has a nice form. While exponential families are useful in their own right, they can also be combined to construct more complex models in a way that is amenable to inference with Markov chain Monte Carlo, as we will see later.

Most of the familiar distributions are exponential families, such as Bernoulli, binomial, Poisson, exponential, beta, gamma, inverse gamma, normal (Gaussian), multivariate normal, log-normal, inverse Gaussian, Dirichlet, and others. On the other hand, the Cauchy distribution and Student's t-distribution are familiar examples that are not exponential families.

The concept of exponential families was developed by E. J. G. Pitman (1897–1993), Bernard Koopman (1900–1981), and Georges Darmois (1888–1960).

| E. J. G. Pitman | Bernard Koopman | Georges Darmois |

2   One-parameter exponential families

A ***one-parameter exponential family*** is a collection of probability distributions indexed by a parameter $\theta \in \Theta$, such that the p.d.f.s/p.m.f.s are of the form

$$p(x|\theta) = \exp\big(\varphi(\theta)t(x) - \kappa(\theta)\big)h(x) \qquad (2.1)$$

for some functions $\varphi(\theta), t(x), \kappa(\theta)$, and $h(x)$. Note that $\kappa(\theta)$ acts a log-normalization constant—for example, in the continuous case, in order to have $\int p(x|\theta)dx = 1$, it must be that

$$\kappa(\theta) = \log \int \exp\big(\varphi(\theta)t(x)\big)h(x)\,dx.$$

Terminology: $t(x)$ is called the ***sufficient statistic function***.

Example 2.1 (Exponential)**.** The $\text{Exp}(\theta)$ distributions form an exponential family, since the p.d.f.s are

$$p(x|\theta) = \text{Exp}(x|\theta) = \theta e^{-\theta x}\mathbb{1}(x > 0) = \exp(-\theta x + \log\theta)\mathbb{1}(x > 0)$$

for $\theta \in \Theta = (0, \infty)$, and we can put this in the form of Equation 2.1 by choosing

$$\begin{aligned}
t(x) &= -x \\
\varphi(\theta) &= \theta \\
\kappa(\theta) &= -\log\theta \\
h(x) &= \mathbb{1}(x > 0).
\end{aligned}$$

Example 2.2 (Poisson)**.** The $\text{Poisson}(\theta)$ distributions form an exponential family, since the p.m.f.s are

$$p(x|\theta) = \text{Poisson}(x|\theta) = \exp(-\theta)\frac{\theta^x}{x!}\mathbb{1}(x \in S) = \exp(x\log\theta - \theta)\mathbb{1}(x \in S)/x!$$

for $\theta \in \Theta = (0, \infty)$, where $S = \{0, 1, 2, \ldots\}$, and we can choose

$$\begin{aligned}
t(x) &= x \\
\varphi(\theta) &= \log\theta \\
\kappa(\theta) &= \theta \\
h(x) &= \mathbb{1}(x \in S)/x!.
\end{aligned}$$

2.1 Natural form

An exponential family is in **natural form** (a.k.a. **canonical form**) if $\varphi(\theta) = \theta$. In this case, θ is called the **natural parameter**.

Example 2.3 (Exponential)**.** Choosing $t(x) = -x$ and $\varphi(\theta) = \theta$ as above puts the family $\mathrm{Exp}(\theta)$ in natural form. If we had instead chosen $t(x) = x$ and $\varphi(\theta) = -\theta$, this would not be in natural form.

Example 2.4 (Poisson)**.** When written as above, the $\mathrm{Poisson}(\theta)$ family is not in natural form. To put it in natural form, we can reparametrize by letting $\eta = \log \theta$. Then

$$p(x|\eta) = \exp(x\eta - e^\eta)\mathbb{1}(x \in S)/x!,$$

for $\eta \in \mathbb{R}$, gives us the same collection of probability distributions, and is in natural form, with natural parameter η.

3 Conjugate priors

Given a family $\{p(x|\theta) : \theta \in \Theta\}$ of generating distributions, a collection of priors $p_\alpha(\theta)$ indexed by $\alpha \in H$ is called a **conjugate prior family** if for any α and any data, the resulting posterior equals $p_{\alpha'}(\theta)$ for some $\alpha' \in H$.

Example 3.1 (Beta-Bernoulli)**.** The collection of $\mathrm{Beta}(\theta|a, b)$ distributions, with $a, b > 0$, is conjugate to $\mathrm{Bernoulli}(\theta)$, since the posterior is $p(\theta|x_{1:n}) = \mathrm{Beta}(\theta \mid a + \sum x_i, \, b + n - \sum x_i)$.

Example 3.2 (Gamma-Exponential)**.** The collection of $\mathrm{Gamma}(\theta|a, b)$ distributions, with $a, b > 0$, is conjugate to $\mathrm{Exp}(\theta)$, since the posterior is $p(\theta|x_{1:n}) = \mathrm{Gamma}(\theta \mid a + n, \, b + \sum x_i)$.

3.1 Existence and uniqueness?

Technically, for any generating family, a conjugate family always exists—namely, the family of all distributions on θ. However, usually when people discuss conjugate priors they mean conjugate families which are computationally-tractable or have a closed-form.

    It is also important to recognize that conjugate priors are usually not unique, so it doesn't make sense to talk about "the conjugate prior" for a given generating family—instead it is more accurate to discuss "a conjugate prior". For instance, suppose $\{p_\alpha(\theta) : \alpha \in H\}$ is a conjugate family, and $g(\theta)$ is any nonnegative function. If $0 < z(\alpha) < \infty$ for all $\alpha \in H$, where $z(\alpha) = \int p_\alpha(\theta)g(\theta)d\theta$, then

$$\big\{ p_\alpha(\theta)g(\theta)/z(\alpha) : \alpha \in H \big\}$$

is also a conjugate family. A useful special case is to take $g(\theta) = \mathbb{1}(\theta \in A)$ for some set A.

3.2 Conjugate priors for exponential families

- Under general conditions, any exponential family has a conjugate prior, with p.d.f.

$$p_{n_0,t_0}(\theta) \propto \exp\left(n_0 t_0 \varphi(\theta) - n_0 \kappa(\theta)\right) \mathbb{1}(\theta \in \Theta)$$

  for the values of $n_0 > 0$ and $t_0 \in \mathbb{R}$ for which this is normalizable.

- However, the normalization constant is not always computationally tractable. Often, when people speak of conjugate priors, it is implied that they mean computationally tractable ones.

- The resulting posterior is $p_{n',t'}(\theta)$ where $n' = n_0 + n$ and

$$t' = \frac{n_0 t_0 + \sum_{i=1}^{n} t(x_i)}{n_0 + n} = \frac{n_0}{n_0 + n} t_0 + \frac{n}{n_0 + n} \frac{1}{n} \sum_{i=1}^{n} t(x_i).$$

  Note that this is a convex combination of t_0 and $\frac{1}{n}\sum t(x_i)$.

- This gives us an insight into the interpretation of the prior parameters t_0 and n_0 that suggests how one might choose them: t_0 represents a prior "guess" at the expected value of $t(x)$, and n_0 represents the prior "number of samples" (roughly speaking, how certain we are about t_0).

4 Multi-parameter exponential families

The generalization to more than one parameter is straightforward. An **exponential family** is a collection of distributions indexed by $\theta \in \Theta$, with p.d.f.s/p.m.f.s of the form

$$p(x|\theta) = \exp\left(\varphi(\theta)^{\mathsf{T}} t(x) - \kappa(\theta)\right) h(x) \tag{4.1}$$

for some vector-valued functions

$$\varphi(\theta) = \begin{pmatrix} \varphi_1(\theta) \\ \vdots \\ \varphi_k(\theta) \end{pmatrix} \quad \text{and} \quad t(x) = \begin{pmatrix} t_1(x) \\ \vdots \\ t_k(x) \end{pmatrix}$$

and some real-valued functions $\kappa(\theta)$ and $h(x)$. Here, v^{T} denotes the transpose of v. As before, $\kappa(\theta)$ is the log-normalization constant, and the content above regarding natural form and conjugate priors applies also to the multi-parameter case.

Example 4.1 (Beta)**.** The Beta(a, b) distributions, with $a, b > 0$, are an exponential family:

$$\text{Beta}(x|a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} \mathbb{1}(0 < x < 1)$$

$$= \exp\left((a-1)\log x + (b-1)\log(1-x) - \log B(a, b)\right) \mathbb{1}(0 < x < 1)$$

which can be put in the form of Equation 4.1 by choosing, for example, $\theta = (a-1, b-1)^{\mathsf{T}}$, $\varphi(\theta) = \theta$, $t(x) = (\log x, \log(1-x))^{\mathsf{T}}$, and $h(x) = \mathbb{1}(0 < x < 1)$.

Example 4.2 (Gamma). The Gamma(a, b) distributions, with $a, b > 0$, are an exponential family:

$$\text{Gamma}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \mathbb{1}(x > 0)$$

$$= \exp\left(-bx + (a - 1)\log x + \log\frac{b^a}{\Gamma(a)}\right) \mathbb{1}(x > 0),$$

and we can choose, for example, $\theta = (a, b)^\mathsf{T}$, $\varphi(\theta) = (-b, a - 1)^\mathsf{T}$, $t(x) = (x, \log x)^\mathsf{T}$, and $h(x) = \mathbb{1}(x > 0)$.

5 Exercises

1. Show that the Bernoulli(θ) distributions form a *one-parameter* exponential family.

2. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Bernoulli(θ), then the distribution of the sum $Z = \sum_{i=1}^n X_i$ has the Binomial(n, θ) distribution. Show that with n fixed, the Binomial(n, θ) distributions form a one-parameter exponential family.

3. Consider the Gamma(a, b) distributions with a fixed. Find a conjugate prior for b.

4. Show that for a certain choice of $t(x)$ and $h(x)$, the Gamma(a, b) distributions are in natural form with natural parameter $\theta = (a, b)^\mathsf{T}$.

5. Section 3.2 presented the general form of the posterior for a one-parameter exponential family with a conjugate prior. Derive this.

6. Suppose $\{p_\alpha(\theta) : \alpha \in H\}$ is a conjugate family for some generator family $\{p(x|\theta) : \theta \in \Theta\}$, let $g(\theta)$ be a nonnegative function, and define $z(\alpha) = \int p_\alpha(\theta) g(\theta) d\theta$. Show that if $0 < z(\alpha) < \infty$ for all $\alpha \in H$, then

$$\{p_\alpha(\theta) g(\theta) / z(\alpha) : \alpha \in H\}$$

is also a conjugate family.

7. Suppose $\{p_\alpha(\theta) : \alpha \in H\}$ is a conjugate family for some generator family. Show that the following is also a conjugate family for the same generator family:

$$\left\{\sum_{i=1}^k \pi_i p_{\alpha_i}(\theta) : \alpha_1, \ldots, \alpha_k \in H, \, \pi \in \Delta_k\right\}$$

where

$$\Delta_k = \left\{\pi \in \mathbb{R}^k : \pi, \ldots, \pi_k \geq 0, \, \sum_{i=1}^k \pi_i = 1\right\}.$$

In other words, finite mixtures of conjugate priors are conjugate priors. This enables one to construct very flexible classes of conjugate priors.

Supplementary material

- Hoff (2009), Section 3.3.

References

- Diaconis, P., & Ylvisaker, D. (1979). Conjugate priors for exponential families. The Annals of Statistics, 7(2), 269-281.

- Orbanz, P., (2009), Functional Conjugacy in Parametric Bayesian Models.

- Hoffman-Jorgensen, J. (1994). Probability with a view towards statistics. CRC Press.

Historical publications

- Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 32, No. 04, pp. 567-579). Cambridge University Press.

- Koopman, B. O. (1936). On distributions admitting a sufficient statistic. Transactions of the American Mathematical Society, 39(3), 399-409.

- Darmois, G. (1935). Sur les lois de probabilite a estimation exhaustive. CR Acad. Sci. Paris, 260, 1265-1266.