Stat260: Bayesian Modeling and Inference **Lecture Date: February 3, 2010**

Conjugate Priors

Lecturer: Michael I. Jordan *Scribe: Steven Troxler*

1 Recap: Dirichlet and Beta Priors

Recall that if $X_1, X_2, ..., X_n$ are i.i.d. draws from a multinomial (n, θ) distribution, then

$$P(X = x|\theta) \propto \theta_1^{\sum_{j=1}^n 1(x_j=\theta_1)} ... \theta_k^{\sum_{j=1}^n 1(x_j=\theta_k)},$$

then a conjugate prior is the Dirichlet distribution with parameter $\alpha \in \mathbb{R}^k$, which has density over the simplex given by

$$P(\theta|\alpha) \propto \theta_1^{\alpha_1-1} ... \theta_k^{\alpha_k-1}.$$

Here we require that for all i, $\alpha_i > 0$. We then have a posterior distribution

$$P(\theta|X, \alpha) \propto \theta_1^{\sum_{j=1}^n 1(x_j=\theta_1)+\alpha_1-1} ... \theta_k^{\sum_{j=1}^n 1(x_j=\theta_k)+\alpha_k-1}.$$

The normalizing constant for the Dirichlet distribution is

$$\frac{1}{B(\alpha)} = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}.$$

We can compute the expectation of θ_i for this distribution using the general fact that $\Gamma(t + 1) = t\Gamma(t)$:

$$\begin{aligned} E[\theta_j|\alpha] &= \int_\triangle \theta_j \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} ... \theta_k^{\alpha_k-1} d\theta \\ &= \frac{\Gamma(\alpha_j + 1)\Gamma(\sum_i \alpha_i)}{\Gamma(\alpha_j)\Gamma(\sum_i \alpha_i + 1)} \int_\triangle \theta_j \frac{\Gamma(\sum_i \alpha_i + 1)}{\Gamma(\alpha_j + 1) \prod_{i\neq j} \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} ... \theta_k^{\alpha_k-1} \\ &= \frac{\Gamma(\alpha_j + 1)\Gamma(\sum_i \alpha_i)}{\Gamma(\alpha_j)\Gamma(\sum_i \alpha_i + 1)} = \frac{\alpha_j}{\sum_i \alpha_i}. \end{aligned}$$

Above, the last line follows because we are integrating a density over the simplex $\triangle$.

A special case is the binomial-beta conjugacy. If $X|\theta$ is distributed as binomial (n, θ), then a conjugate prior is the beta family of distributions, defined by the density

$$p(\theta|\alpha_1, \alpha_2) \propto \theta^{\alpha_1-1}(1 - \theta)^{\alpha_2-1}.$$

The work above shows that

$$E[\theta|\alpha_1, \alpha_2] = \frac{\alpha_1}{\alpha_1 + \alpha_2}.$$

As a comparison of the $\alpha_1 = \alpha_2 = 1/2$ and $\alpha_1 = \alpha_2 = 2$ cases in Figure 1 suggest, when the parameters are equal the prior mean of the beta is 1/2, and the prior variance decreases as the parameters grow. These observations carry over to the more general Dirichlet distribution, which becomes more concentrated as $\sum_i \alpha_i$ becomes large, so that $\text{Var}(\theta|\alpha) \downarrow 0$ as $\sum_i \alpha_i \uparrow \infty$.

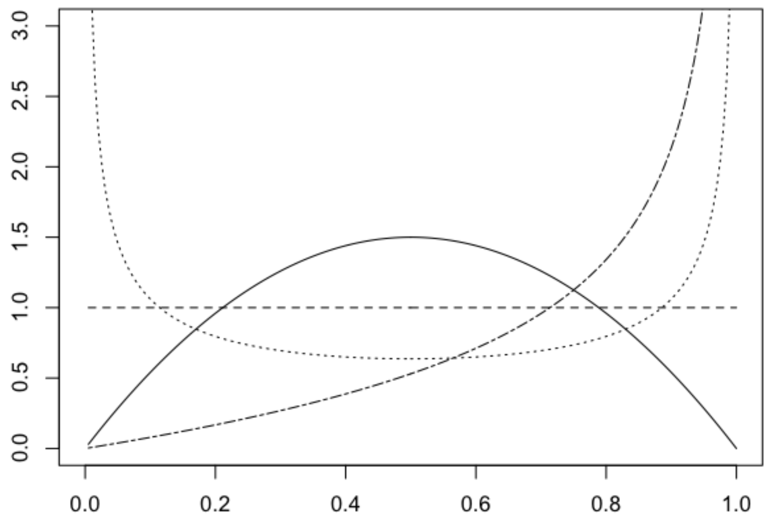

Figure 1: A plot of several beta densities. The flat line corresponds to $\alpha_1 = \alpha_2 = 1$, which gives a uniform distribution. The other cases are $\alpha_1 = \alpha_2 = 1/2$, the dotted line, $\alpha_1 = \alpha_2 = 2$, the solid line, and $\alpha_1 = 2, \alpha_2 = 1/2$, the dot-dash line.

2 Multinomial Dirichlet Conjugacy

It is clear that if $X|\theta$ is distributed as multinomial (n, θ) and $\theta|\alpha$ is distributed as Dirichlet with parameter α, then $\theta|\alpha, X$ will have density

$$p(\theta|x, \alpha) \propto \theta_1^{\alpha_1 + \sum_{j=1}^n 1(x_j=1)-1} ... \theta_k^{\alpha_k + \sum_{j=1}^n 1(x_j=k)-1}.$$

The normalizing constant is given by

$$\frac{\Gamma(\sum_i \alpha_i + n)}{\prod_i \Gamma(\alpha_i + \sum_{j=1}^n 1(x_j = i))}.$$

The posterior mean is given by

$$E[\theta_i|x, \alpha] = \frac{\alpha_i + \sum_{j=1}^n 1(x_j = i)}{n + \sum_{l=1}^k \alpha_l} = \kappa \frac{\alpha_i}{\sum_{l=1}^k \alpha_l} + (1-\kappa)\bar{x}_i,$$

where $\bar{x}_i = \frac{1}{n}\sum_{j=1}^n 1(x_j = i)$ is the maximum likelihood estimator (exercise: check this) and $\kappa = \frac{\sum_l \alpha_l}{n + \sum_l \alpha_l} \in (0, 1)$.

Several features of this posterior mean are worth observing. First, it is a convex combination of the maximum likelihood estimate and the prior mean. For this reason, it is sometimes called a *shrinkage* estimator, especially when the prior mean takes some central value such as setting all the parameters equal to $1/k$. Second, the convex combination is determined by κ, which decreases to 0 as $n \uparrow \infty$. For this reason, the posterior mean is asymptotically optimal, since for large n it behaves like the maximum likelihood estimator. Both these features, we will see, are general features of the conjugate priors to exponential families.

For small n, the degree to which shrinkage takes place is determined by $\sum_l \alpha_l$. In other words, the bigger $\sum_l \alpha_l$, the less spread out our prior is and therefore the more confidence we have in the prior mean before looking at the data. When $\sum_l \alpha_l$ is large compared to n, the prior will tend to dominate the data.

3 Poisson Gamma Conjugacy

Suppose now that $X|\theta$ has a Poisson(θ) distribution, or more generally, that $X_1, ..., X_n|\theta$ is an iid sample from a Poisson(θ) distribution. Then X has conditional density

$$p(x|\theta) = \prod_{j=1}^{n} \frac{\theta^{x_j} e^{-\theta}}{x_j!} \propto \theta^{\sum_j x_j} e^{-n\theta}.$$

A conjugate prior is the gamma (α_1, α_2) distribution, with density

$$p(\theta|\alpha_1, \alpha_2) \propto \theta^{\alpha_1 - 1} e^{-\alpha_2 \theta},$$

where the normalizing constant is $\alpha_2^{\alpha_1}/\Gamma(\alpha_1)$. The expectation of this distribution may be calculated using a method similar to that we used for the Dirichlet distribution, and we find that $E[\theta|\alpha_1, \alpha_2] = \frac{\alpha_1}{\alpha_2}$.

The posterior distribution has density

$$P(\theta|x, \alpha) \propto \theta^{\sum_j x_j + \alpha_1 - 1} e^{-(\alpha_2 + n)\theta},$$

so that

$$E[\theta|x, \alpha] = \frac{\sum_j x_j + \alpha_1}{n + \alpha_2} = \kappa \frac{\alpha_1}{\alpha_2} + (1 - \kappa) \frac{\sum_j x_j}{n},$$

where $\kappa = \alpha_1/(\alpha_2 + n)$. Again, we see that this is a convex combination of the prior mean and maximum likelihood estimate, and that it is asymptotically equivalent to the MLE.

4 Conjugacy for General Exponential Families

In general, an exponential family is one with a density (typically with respect to Lebesgue measure or counting measure) given by

$$p(x|\eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\},$$

so that if $X_1, X_2, ..., X_n$ is an iid sample from the same distribution, conditional on η, the sample has conditional density of the form

$$p(x|\eta) = \prod_j [h(x_j)] \exp\{\eta^T \sum_j T(x_j) - nA(\eta)\}.$$

We define a conjugate prior for this exponential family by taking

$$p(\eta|\tau, n_0) = H(\tau, n_0) \exp\{\tau^T \eta - n_0 A(\eta)\},$$

another exponential family. In the posterior distribution, the hyperparameter τ is updated to $\tau + \sum_j T(x_j)$, while the hyperparameter n_0 is updated to $n + n_0$.

Set $\mu = \mu(\eta) = E[T(x)|\eta]$. From the theory of exponential families, we know $\mu = \triangledown_\eta A(\eta)$, where $\triangledown$ denotes the gradient. We want to treat μ like a parameter, and find its expected value with respect to the prior and posterior. We will make use of Green's theorem to do so. First, we note that

$$E[\mu|\tau, n_0] = E[\triangledown_\eta A(\eta)|\tau, n_0]$$

and, by direct compuatation,

$$\triangledown p(\eta|\tau, n_0) = p(\eta|\tau, n_0)(\tau - n_0 \triangledown_\eta A(\eta)).$$

Now, since $p(\eta|\tau, n_0)$ is a density, hence zero at the edges of $\mathbb{R}^p$, Green's theorem ensures that

$$\int_{\mathbb{R}^p} p(\eta|\tau, n_0)(\tau - n_0 \nabla_\eta A(\eta))d\eta = \int_{\mathbb{R}^p} \nabla p(\eta|\tau, n_0)d\eta = 0.$$

Since the term on the left is just $\tau - n_0 E[\nabla_\eta A(\eta)|\tau, n_0]$, this tells us that

$$E[\mu|\tau, n_0] = E[\nabla_\eta A(\eta)|\tau, n_0] = \frac{\tau}{n_0},$$

and hence also that

$$E[\mu|\tau, n_0] = \frac{\tau + \sum_j T(x_j)}{n + n_0} = \kappa \frac{\tau}{n_0} + (1-\kappa)\frac{\sum_j T(x_j)}{n},$$

with $\kappa = \frac{n_0}{n_0 + n}$.

Remark: Diaconis and Ylvisaker prove that, under mild conditions, the converse holds: if the posterior mean is always a convex combination of the MLE and prior mean, then we are working in an exponential family.

5 Gaussian and Conjugate Prior

The Gaussian distribution with parameters μ and σ^2 has density

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

Conjugate priors for the Gaussian distribution are easy to find if one of μ or σ^2 are known, so that we only have to worry about one parameter. It is left for the reader, for instance, to check (via completion of the square) that a normal distribution provides a conjugate prior for μ if σ^2 is fixed. If μ is fixed, a conjugate prior for σ^2 is the inverse gamma.

We conclude this lecture by defining the inverse gamma density, and will pick up here next lecture. Suppose y has a gamma (α, β) distribution, so that

$$p(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y},$$

and let $z = 1/y$, so that $y = 1/z$ and $dy/dz = -1/z^2$. By the change of variables formula,

$$p(z|\alpha, \beta) = p(y(z)|\alpha, \beta)\left|\frac{dy}{dz}\right| = \frac{\beta^\alpha}{\Gamma(\alpha)} y(z)^{\alpha-1} e^{-\beta y(z)} \frac{1}{z^2} = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} e^{-\beta/z}.$$

This density defines the inverse gamma distribution.