MAS3301 Bayesian Statistics

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 2, 2008-9

11 Conjugate Priors IV: The Dirichlet distribution and multinomial observations

11.1 The Dirichlet distribution

The Dirichlet distribution is a distribution for a set of quantities $\theta_1, \dots, \theta_m$ where $\theta_i \geq 0$ and $\sum_{i=1}^m \theta_i = 1$. An obvious application is to a set of probabilities for a partition (i.e. for an exhaustive set of mutually exclusive events).

The probability density function is

$$f(\theta_1, \dots, \theta_m) = \frac{\Gamma(A)}{\prod_{i=1}^m \Gamma(a_i)} \prod_{i=1}^m \theta_i^{a_i - 1}$$

where $A = \sum_{i=1}^m a_i$ and $a_1, \dots, a_m$ are parameters with $a_i > 0$ for $i = 1, \dots, m$.

Clearly, if $m = 2$, we obtain a beta(a_1, a_2) distribution as a special case.

The mean of θ_j is

$$\mathrm{E}(\theta_j) = \frac{a_j}{A}$$

the variance of θ_j is

$$\mathrm{var}(\theta_j) = \frac{a_j}{A(A+1)} - \frac{a_j^2}{A^2(A+1)}$$

and the covariance of θ_j and θ_k, where $j \neq k$, is

$$\mathrm{covar}(\theta_j, \theta_k) = -\frac{a_j a_k}{A^2(A+1)}.$$

Also the marginal distribution of θ_j is beta$(a_j,\ A - a_j)$.

Note that the space of the parameters $\theta_1, \dots, \theta_m$ has only $m - 1$ dimensions because of the constraint $\sum_{i=1}^m \theta_i = 1$, so that, for example, $\theta_m = 1 - \sum_{i=1}^{m-1} \theta_i$. Therefore, when we integrate over this space, the integration has only $m - 1$ dimensions.

Proof (mean)

The mean is

$$\begin{aligned}
\mathrm{E}(\theta_j) &= \int \cdots \int \theta_j \frac{\Gamma(A)}{\prod_{i=1}^m \Gamma(a_i)} \prod_{i=1}^m \theta_i^{a_i - 1}\ d\theta_1 \dots d\theta_{m-1} \\
&= \frac{\Gamma(A)}{\Gamma(A+1)} \frac{\Gamma(a_j+1)}{\Gamma(a_j)} \int \cdots \int \frac{\Gamma(A+1)}{\prod_{i=1}^m \Gamma(a_i')} \prod_{i=1}^m \theta_i^{a_i' - 1}\ d\theta_1 \dots d\theta_{m-1} \\
&= \frac{\Gamma(A)}{\Gamma(A+1)} \frac{\Gamma(a_j+1)}{\Gamma(a_j)} = \frac{a_j}{A}
\end{aligned}$$

where $a_i' = a_i$ when $i \neq j$ and $a_j' = a_j + 1$.

Proof (variance)

Similarly

$$\mathrm{E}(\theta_j^2) = \frac{\Gamma(A)}{\Gamma(A+2)} \frac{\Gamma(a_j+2)}{\Gamma(a_j)} = \frac{(a_j+1)a_j}{(A+1)A}$$

so

$$\mathrm{var}(\theta_j) = \frac{(a_j+1)a_j}{(A+1)A} - \left(\frac{a_j}{A}\right)^2 = \frac{a_j}{A(A+1)} - \frac{a_j^2}{A^2(A+1)}$$

Proof (covariance)

Also

$$\mathrm{E}(\theta_j\theta_k) = \frac{\Gamma(A)}{\Gamma(A+2)} \frac{\Gamma(a_j+1)}{\Gamma(a_j)} \frac{\Gamma(a_k+1)}{\Gamma(a_k)} = \frac{a_j a_k}{(A+1)A}$$

so

$$\mathrm{covar}(\theta_j, \theta_k) = \frac{a_j a_k}{(A+1)A} - \frac{a_j}{A}\frac{a_k}{A} = -\frac{a_j a_k}{A^2(A+1)}$$

Proof (marginal)

We can write the joint density of $\theta_1, \ldots, \theta_m$ as

$$f_1(\theta_1) f_2(\theta_2 \mid \theta_1) f_3(\theta_3 \mid \theta_1, \theta_2) \cdots f_{m-1}(\theta_{m-1} \mid \theta_1, \ldots, \theta_{m-2}).$$

(We do not need to include a final term in this for θ_m because θ_m is fixed once $\theta_1, \ldots, \theta_{m-1}$ are fixed).

In fact we can write the joint density as

$$\frac{\Gamma(A)}{\Gamma(a_1)\Gamma(A-a_1)} \theta_1^{a_1-1}(1-\theta_1)^{A-a_1-1} \times \frac{\Gamma(A-a_1)}{\Gamma(a_2)\Gamma(A-a_1-a_2)} \frac{\theta_2^{a_2-1}(1-\theta_1-\theta_2)^{A-a_1-a_2-1}}{(1-\theta_1)^{A-a_1-1}}$$

$$\times \cdots \times \frac{\Gamma(A-a_1-\cdots-a_{m-2})}{\Gamma(a_{m-1})\Gamma(A-a_1-\cdots-a_{m-1})} \frac{\theta_{m-1}^{a_{m-1}-1}\theta_m^{a_m-1}}{(1-\theta_1-\cdots\theta_{m-2})^{a_{m-1}+a_m-1}}.$$

A bit of cancelling shows that this simplifies to the correct Dirichlet density.

Thus we can see that the marginal distribution of θ_1 is a beta$(a_1,\ A - a_1)$ distribution and similarly that the marginal distribution of θ_j is a beta$(a_j,\ A-a_j)$ distribution. We can also deduce the distribution of a subset of $\theta_1, \ldots, \theta_m$. For example if $\tilde{\theta}_3 = 1-\theta_1-\theta_2-\theta_3$, then the distribution of $\theta_1, \theta_2, \theta_3, \tilde{\theta}_3$ is Dirichlet$(a_1,\ a_2,\ a_3,\ \tilde{a}_3)$ where $\tilde{a}_3 = A - a_1 - a_2 - a_3$.

11.2 Multinomial observations

11.2.1 Model

Suppose that we will observe $X_1, \ldots, X_m$ where these are the frequencies for categories $1, \ldots, m$, the total $N = \sum_{i=1}^m X_i$ is fixed and the probabilities for these categories are $\theta_1, \ldots, \theta_m$ where $\sum_{i=1}^m \theta_i = 1$. Then, given θ, where $\theta = (\theta_1, \ldots, \theta_m)^T$, the distribution of $X_1, \ldots, X_m$ is multinomial with

$$\Pr(X_1 = x_1, \ldots, X_m = x_m) = \frac{N!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m \theta_i^{x_i}.$$

Notice that, with $m = 2$, this is just a binomial(N, θ_1) distribution.

Then the likelihood is

$$\begin{aligned} L(\theta;\ x) &= \frac{N!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m \theta_i^{x_i} \\ &\propto \prod_{i=1}^m \theta_i^{x_i}. \end{aligned}$$

The conjugate prior is a *Dirichlet* distribution which has a pdf proportional to

$$\prod_{i=1}^m \theta_i^{a_i - 1}.$$

The posterior pdf is proportional to

$$\prod_{i=1}^m \theta_i^{a_i-1} \times \prod_{i=1}^m \theta_i^{x_i} = \prod_{i=1}^m \theta_i^{a_i+x_i-1}.$$

This is proportional to the pdf of a Dirichlet distribution with parameters $a_1 + x_1, a_2 + x_2, \ldots a_m + x_m$.

11.2.2 Example

In a survey 1000 English voters are asked to say for which party they would vote if there were a general election next week. The choices offered were 1: Labour, 2: Liberal, 3: Conservative, 4: Other, 5: None, 6: Undecided. We assume that the population is large enough so that the responses may be considered independent given the true underlying proportions. Let $\theta_1, \ldots, \theta_6$ be the probabilities that a randomly selected voter would give each of the responses. Our prior distribution for $\theta_1, \ldots, \theta_6$ is a Dirichlet(5, 3, 5, 1, 2, 4) distribution.

This gives the following summary of the prior distribution.

Response	a_i	Prior mean	Prior var.	Prior sd.
Labour	5	0.25	0.008929	0.09449
Liberal	3	0.15	0.006071	0.07792
Conservative	5	0.25	0.008929	0.09449
Other	1	0.05	0.002262	0.04756
None	2	0.10	0.004286	0.06547
Undecided	4	0.20	0.007619	0.08729
Total	20	1.00		

Suppose our observed data are as follows.

Labour	Liberal	Conservative	Other	None	Undecided
256	131	266	38	114	195

Then we can summarise the posterior distribution as follows.

Response	$a_i + x_i$	Posterior mean	Posterior var.	Posterior sd.
Labour	261	0.2559	0.0001865	0.01366
Liberal	134	0.1314	0.0001118	0.01057
Conservative	271	0.2657	0.0001911	0.01382
Other	39	0.0382	0.0000360	0.00600
None	116	0.1137	0.0000987	0.00994
Undecided	199	0.1951	0.0001538	0.01240
Total	1020	1.0000		

12 Sufficiency

12.1 Introduction

Consider the following problem. We are going to observe two random variables X_1, X_2. In each case, given the value of μ, we have

$$X_i \mid \mu \sim N(\mu,\ V)$$

where the variance V is known but we wish to learn about the value of μ. Further, given μ, the two variables X_1, X_2 are independent.

The likelihood comes from the joint pdf of X_1, X_2 but an exactly equivalent observation would be Y_1, Y_2 where

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= X_1 - X_2 \end{aligned}$$

It is easily seen that

$$\begin{aligned} Y_1 &\sim N(2\mu,\ 2V) \\ Y_2 &\sim N(0,\ 2V) \end{aligned}$$

and that Y_1 and Y_2 are independent. Therefore Y_2 does not depend on μ and its value can not tell us anything about μ. On the other hand the value of Y_1 tells us everything which we can learn from the data about μ. We say that Y_1 is *sufficient* for μ and Y_2 is *ancillary* for μ.

12.2 Definition

Suppose we have an unknown (e.g. a parameter) θ and we will observe data Y. The density (or probability) of Y given θ is $f_{Y|\theta}(y \mid \theta)$ and this gives us the likelihood, $L(\theta;\ y)$. Suppose we have a statistic $T(Y)$, with value t.

Since, once we know Y, we can calculate T, can always write

$$f_{Y|\theta}(y \mid \theta) = f_{Y,T|\theta}(y, t \mid \theta) = f_{T|\theta}(t \mid \theta) f_{Y|t,\theta}(y \mid t, \theta).$$

In some cases $f_{Y|t,\theta}(y \mid t, \theta)$ does not depend on θ so $f_{Y|t,\theta}(y \mid t, \theta) = f_{Y|t}(y \mid t)$. In this case

$$f_{Y|\theta}(y \mid \theta) = f_{T|\theta}(t \mid \theta) f_{Y|t}(y \mid t). \tag{9}$$

In such a case we say that $T(Y)$ is a *sufficient statistic* for θ given Y. Often we simply say that T is *sufficient* for θ.

12.3 Factorisation theorem

Another way to express (9) is to say that T is sufficient for θ if and only if there are functions $g,\ h$ such that

$$f_{Y|\theta}(y \mid \theta) = g(\theta, t) h(y) \tag{10}$$

where $h(y)$ does not depend on θ.

This is known as Neyman's factorisation theorem.

Proof: If T is sufficient for θ then we can write $g(\theta, t) = f_{T|\theta}(t \mid \theta)$ and $h(y) = f_{Y|t}(y \mid t)$.

To prove the converse we start by integrating (or summing) (10) over all values of y where $T(y) = t$. This gives

$$f_{T|\theta}(t \mid \theta) = g(\theta, t) H(t)$$

for some function $H(t)$. This gives us

$$g(\theta, t) = \frac{f_{T|\theta}(t \mid \theta)}{H(t)}$$

which we substitute in (10) to obtain

$$f_{Y|\theta}(y \mid \theta) = \frac{f_{T|\theta}(t \mid \theta)h(y)}{H(t)}.$$

Now

$$f_{Y|t,\theta}(y \mid t, \theta) = \frac{f_{Y,T|\theta}(y, t \mid \theta)}{f_{T|\theta}(t \mid \theta)} = \frac{f_{Y|\theta}(y \mid \theta)}{f_{T|\theta}(t \mid \theta)}$$

so

$$f_{Y|t,\theta}(y \mid t, \theta) = \frac{h(y)}{H(t)}.$$

The right hand side of this equation does not depend on θ so the theorem is proved.

12.4 Sufficiency principle

From (9) we can see that, if T is sufficient for θ, then the likelihood for θ from y is proportional to the likelihood for θ from t. Therefore, instead of using the likelihood for the full data we can use the likelihood based simply on the distribution of T.

12.5 Examples

12.5.1 Poisson

Suppose that we observe random variables $Y_1, \dots, Y_n$ where, given the value of the parameter λ, Y_i is independent of Y_j for $i \neq j$ and $Y_i \sim \text{Poisson}(\lambda)$ for $i = 1, \dots, n$.
Then the likelihood is

$$\begin{aligned} L(\lambda;\ y) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} &= e^{-n\lambda}\lambda^{S} \prod_{i=1}^{n} \frac{1}{y_i!} \\ &= g(\lambda, S)h(y) \end{aligned}$$

where $S = \sum_{i=1}^{n} y_i,\ g(\lambda, S) = e^{-n\lambda}\lambda^S$ and $h(y) = \prod_{i=1}^{n} \frac{1}{y_i!}$. So S is sufficient for λ.
Furthermore $S \sim \text{Poisson}(n\lambda)$ so an equivalent likelihood is

$$L_S(\lambda; y) = \frac{e^{-n\lambda}(n\lambda)^S}{S!} \propto e^{-n\lambda}\lambda^S.$$

12.5.2 Normal

Suppose that we observe random variables $Y_1, \ldots, Y_n$ where, given the value of the parameters $\mu,\ \sigma^2$, Y_i is independent of Y_j for $i \neq j$ and $Y_i \sim N(\mu,\ \sigma^2)$ for $i = 1, \ldots, n$.
Here the parameter is $\theta = (\mu, \sigma^2)^T$.
The likelihood is

$$\begin{aligned}
L(\mu, \sigma^2;\ y) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)^2\right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right]\right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[S + n(\bar{y} - \mu)^2\right]\right\} \\
&= g(\theta, T)h(y)
\end{aligned}$$

where $h(y) = 1$, $T = (\bar{y}, S)^T$,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{and} \qquad S = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Hence $\bar{y}$ and S are sufficient for μ and σ^2.
Furthermore, in the case where σ^2 is known, $\bar{y}$ is sufficient for μ since

$$\begin{aligned}
L(\mu;\ y) &= \exp\left\{-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2\right\}(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{S}{2\sigma^2}\right\} \\
&= g(\mu, \bar{y})h(y)
\end{aligned}$$

with

$$h(y) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{S}{2\sigma^2}\right\}.$$