Bayesian Statistics III/IV — Handout on multinomial inference and the Dirichlet distribution, leading to the Dirichlet process and Bayesian non-parametric inference

Peter Craig
April 16 2009

1 Multinomial inference

Throughout this first section of the handout, we are dealing with Bayesian inference for a multinomial population which has \( M \) categories labelled \( V_1, \ldots, V_M \).

1.1 Likelihood function

We imagine that the population is (effectively) infinite and that proportion \( \Theta_i \) of the population is in category \( V_i \) so that \( \Theta_1 + \cdots + \Theta_M = 1 \) and, if we take a random item from the population, \( \Theta_i \) is the probability that it is from category \( V_i \).

If we now take a random sample of size \( n \) from the population, we can make inferences about \( \Theta = (\Theta_1, \ldots, \Theta_M) \). As in the binomial case, we simply multiply the individual probabilities of the individual observations to obtain the likelihood if we see the order of the observations. Then the likelihood function is \( l(\Theta) = \Theta_{k_1} \cdots \Theta_{k_M} \) where \( k_i \) is the number of observations which fall in category \( V_i \) and \( k_1 + \cdots + k_M = n \). If we don’t see the order of the observations, we need also to count the number of different possible orderings using the multinomial coefficient

\[
\binom{n}{k_1, k_2, \ldots, k_M} = \frac{n!}{k_1! \cdots k_M!},
\]

and the likelihood function becomes

\[
l(\Theta) = \left( \binom{n}{k_1, k_2, \ldots, k_M} \right) \Theta_{k_1} \cdots \Theta_{k_M} \]

However, from the perspective of Bayesian inference, the multinomial coefficient is irrelevant as it does not change the posterior pdf (probability density function) which is obtained by re-normalising the product of prior pdf and likelihood to integrate to 1.

1.2 Conjugate prior

It’s now fairly obvious that if we take a prior pdf of the form

\[
p(\Theta) \propto \Theta_{a_1} \cdots \Theta_{a_M} \quad \text{for } a_1 \geq 0 \text{ and } a_1 + \cdots + a_M = 1 \tag{1}\]

we have a conjugate prior family, as then

\[
p(\Theta \mid \text{data}) \propto l(\Theta)p(\Theta) \propto \Theta_{a_1+k_1} \cdots \Theta_{a_M+k_M} \tag{2}\]

which is of the same form but where \( a_i \) has updated to \( a_i + k_i \).

In one sense this tells us everything we need to know about simple Bayesian inference for multinomial populations. However, to exploit (2), we need to be able to choose \( a_1, \ldots, a_M \) for our prior distribution and to interpret what the numbers \( a_1 + k_1, \ldots, a_M + k_M \) tell us about our posterior distribution.

2 Dirichlet distributions

The distributions defined in (1) are known as Dirichlet distributions and it is easy to see that they are beta-distributions when \( M = 2 \) since then \( \Theta_1 = 1 - \Theta_1 \). We want to know much more; the easiest way to understand Dirichlet distributions is to look at a special way of constructing \( \Theta \) from independent gamma-distributed quantities.

2.1 Construction

Let \( W_i \sim \Gamma(\alpha_i, 1) \) be independent for \( i = 1, \ldots, M \) and set \( \Theta_i = W_i/T \) where \( T = W_1 + \cdots + W_M \). Note that this imposes the constraint that \( \Theta_1 + \cdots + \Theta_M = 1 \) and that \( \Theta_i \geq 0 \) since \( W_i \) cannot be negative.

In the next paragraphs, we shall see that the pdf for \( \Theta \) is given by (1) with \( a_i = \alpha_i - 1 \).

Exploiting the definition of the gamma-distribution and independence of \( W_1, \ldots, W_M \), we have the following pdf for \( W = (W_1, \ldots, W_M) \):

\[
p(w) = \prod_i \frac{\alpha_i - 1}{\Gamma(\alpha_i)} e^{-w_i} \]

where \( \Gamma(\alpha_i) \) is the gamma function.
We can now obtain the pdf for \( \Theta \) by a multivariable change of variable (including Jacobian) but we must be careful since \( W \) is \( M \)-dimensional and \( \Theta \) is effectively only \((M-1)\)-dimensional due to the constraint that \( \Theta_1 + \cdots + \Theta_M = 1 \). The theory of how to to a multivariable change of variable only applies if both sides of the change are the same dimension. Therefore, we change instead from \( W_1, \ldots, W_M \) to \( \Theta_1, \ldots, \Theta_{M-1} \), noting that this is bijective since we can write \( W_i = T\Theta_i \) for \( i < M \) and \( W_M = T(1 - \Theta_1 - \cdots - \Theta_{M-1}) \).

We need the Jacobian: For \( i < M \), \( \partial W_i / \partial \Theta_j = T \delta_{ij} \) and \( \partial W_i / \partial T = \Theta_i \). \( \partial W_M / \partial \Theta_j = -T \) and \( \partial W_M / \partial T = (1 - \Theta_1 - \cdots - \Theta_{M-1}) \) which we shall continue to write as \( \Theta_M \). Hence the determinant is

\[
\begin{vmatrix}
T & \Theta_1 \\
T & \Theta_2 \\
\vdots & \vdots \\
-T & -T & \cdots & -T & \Theta_M \\
0 & 0 & \ldots & 0 & 1
\end{vmatrix}
\]

Adding all other rows to the final row results in

\[
\begin{vmatrix}
T & \Theta_1 \\
T & \Theta_2 \\
\vdots & \vdots \\
T & \Theta_{M-1} \\
0 & 0 & \ldots & 0 & 1
\end{vmatrix}
\]

without changing the determinant. Since all entries below the diagonal are zero, the determinant is the product of the diagonal entries\(^4\): \( T^{M-1} \).

Hence, substituting in \( p(w) \) and multiplying by \( t^{M-1} \), we have

\[
p(\theta_1, \ldots, \theta_{M-1}, t) = t^{M-1} \prod_{i=1}^M (\theta_i t)^{\alpha_i - 1} e^{-\theta_i t}
\]

\[
= t^{M-1+\alpha_1+\cdots+\alpha_{M-1}} e^{-(\theta_1+\cdots+\theta_{M-1})t} \prod_i \frac{\theta_i^{\alpha_i-1}}{\Gamma(\alpha_i)}
\]

\[
= t^{A-1} e^{-t} \prod_i \frac{\theta_i^{\alpha_i-1}}{\Gamma(\alpha_i)}
\]

where \( A = \alpha_1 + \cdots + \alpha_M \).

Now we integrate with respect to \( t \) in order to obtain \( p(\theta_1, \ldots, \theta_{M-1}) \). The terms involving \( t \) are

\[
t^{A-1} e^{-t}
\]

which would be the pdf of \( \Gamma(A, 1) \) except that the normalising constant is missing and so the integral is \( \Gamma(A) \) which gives

\[
p(\theta_1, \ldots, \theta_{M-1}) = \frac{\Gamma(A)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_M)} \theta_1^{\alpha_1-1} \cdots \theta_{M-1}^{\alpha_{M-1}-1}
\]

(3)

so that \( \alpha_i = \alpha_i - 1 \) in (1) as claimed earlier. Note that we now know the missing normalising constant in (1).

\( \theta_M \) is missing from the list of variables in the pdf on the left hand side of (3) but \( \Theta_M \) is determined by \( \Theta_1, \ldots, \Theta_{M-1} \) and so (3) is also \( p(\theta_1, \ldots, \theta_M) \), sometimes denoted by \( p(\theta) \).

We write \( \Theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_M) \) (or Dirichlet(\( \alpha \)) for short) and \( \alpha_1, \ldots, \alpha_M \) are called the “degrees of freedom” of the Dirichlet distribution.

2.2 Properties

We now explore some properties of the Dirichlet distribution which are useful both for performing calculations and for understanding the material on Bayesian non-parametric inference at the end of the handout.

2.2.1 Conglomeration

Suppose we form \( W_* \) by summing \( W_1 \) and \( W_2 \). One of the key properties of the gamma-distribution is that you can add independent gamma-distributed quanti-

---

\(^4\)The individual terms being summed in a determinant are obtained by multiplying \( M \) entries in the matrix chosen so that each row contributes a single entry and each column contributes a single entry. The only non-zero entry in the first column is the first diagonal entry. The remaining entries cannot come from either the first row or first column and so the only remaining non-zero entry in the second column is the second diagonal entry. By induction the only non-zero contribution to the determinant is the product of the diagonal entries.
ties and obtain another gamma-distribution, provided the two original distributions have the same rate parameter.

The mgf (moment generating function) of $\Gamma(\alpha, 1)$ is $\Gamma(\alpha, 1) = 1/(1-u)^\alpha$. Note that all our $W_i$ have 1 as the rate parameter. Thus the mgf of $W = W_1 + W_2$ is

$$\frac{1}{(1-u)^{\alpha_1}} \frac{1}{(1-u)^{\alpha_2}} = \frac{1}{(1-u)^{\alpha_1+\alpha_2}}$$

which is the mgf of $\Gamma(\alpha_1 + \alpha_2, 1)$. Note that $W_1, W_3, \ldots, W_M$ are independent since $W_1$ is computed from $W_1$ and $W_2$ which were independent of $W_3, \ldots, W_M$. Consequently the same construction based on $W_1, W_3, \ldots, W_M$ leads to a Dirichlet distribution for $\Theta_1, \Theta_2, \ldots, \Theta_M$ with $\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_M$ degrees of freedom where $\theta_i$ is unchanged for $i > 2$ and $\Theta_i = W_i/(W_1 + W_3 + \cdots + W_M) = (W_i + W_2)/(W_1 + \cdots + W_M) = \Theta_1 + \Theta_2$.

Note that the total degrees of freedom $A = \alpha_1 + \cdots + \alpha_M$ is unchanged by conglomeration. Symmetry implies that the preceding argument works for combining any two categories and clearly we can apply induction to combine any number of categories in a similar manner.

### 2.2.2 Marginal distributions

By combining all categories other than $V_i$, we find that $(\theta_i, 1 - \theta_i) \sim Dirichlet(\alpha_i, A - \alpha_i)$ and so the pdf of $\Theta_i$ is

$$\frac{\Gamma(A)}{\Gamma(\alpha_i) \Gamma(A - \alpha_i)} \theta_i^{\alpha_i} (1 - \theta_i)^{A - \alpha_i}$$

i.e. $\Theta_i \sim \beta(\alpha_i, A - \alpha_i)$.

Consequently, from familiar properties of the beta-distribution, we can compute

$$E[\Theta_i] = \frac{\alpha_i}{\alpha_i + (A - \alpha_i)} = \frac{\alpha_i}{A}$$

and

$$Var[\Theta_i] = \frac{\alpha_i(A - \alpha_i)}{A^2(A + 1)} = \frac{E[\Theta_i](1 - E[\Theta_i])}{A + 1}$$

$\hat{\text{You can calculate this yourself for } W \sim \Gamma(\alpha, 1) \text{ using } m(u) = E[e^{uW}] = \int_0^\infty e^{wu}u^{\alpha-1}e^{-w} dw}^5$

### 3 Multinomial inference re-visited

We have seen that if we take $\Theta \sim Dirichlet(\alpha)$ as our prior distribution, then our posterior distribution is $\Theta \mid \text{data} \sim Dirichlet(\alpha + k)$ which has total degrees of freedom $A + (k_1 + \cdots + k_M) = A + n$.

We can choose our prior distribution as follows. Write down our (subjectively specified) prior mean $E(\Theta_i)$ for the proportion of the population in each category $V_i$. This determines $\alpha_i/A$ for $i = 1, \ldots, M$. To complete the task, we need to specify a value for $A$ and we can then determine the value of each $\alpha_i$. However, we have seen in the previous section that the role of $A$ is to control the prior variance of $\Theta_i$ and so we choose $A$ to give an appropriate degree of prior uncertainty (prior standard deviation) about each $\Theta_i$.

Note that, if we know the population distribution, $\Theta_i$ is simply the probability a randomly chosen value from the population falls in $V_i$. Thus, when specifying my prior expectation for $\Theta_i$, I am simply specifying my prior probability that a randomly chosen value from the population lies in category $V_i$.

We can interpret the posterior distribution through the same process. Our posterior mean for the proportion of the population in each category is

$$E(\Theta_i \mid \text{data}) = \frac{k_i}{A + n} = \frac{A \cdot \alpha_i}{A + n A} + \frac{n \cdot k_i}{A + n} = wE(\Theta_i) + (1 - w) \hat{\theta}_i$$

which is a weighted average of the prior mean for the proportion of population in $V_i$ and the sample proportion $\hat{\theta}_i = k_i/n$ in $V_i$ where the weight $w = A/(A + n)$ represents the fraction of evidence coming from the prior.

The posterior standard deviation of the proportion of population in $V_i$ is then

$$\sqrt{\frac{E[\Theta_i \mid \text{data}](1 - E[\Theta_i \mid \text{data})]}{A + n + 1}}$$

so that posterior uncertainty about each population proportion decays at rate $1/\sqrt{A + n + 1}$ as the sample size increases. Note that it is the total amount of evidence $A + n$ which appears here rather than just the sample size $n$.

### 4 Bayesian non-parametrics

Suppose now that we want to carry out non-parametric inference for a population of numerical values in some (known) interval $(L, R)$.
Essentially we need a way to have prior and posterior probabilities for all possible population distributions.

4.1 Categorisation

The easy way to proceed is to divide the interval up into a number of sub-intervals and learn about the proportion of the population in each interval. Let \( x_0 < x_1 < \cdots < x_M \) where \( x_0 = L \) and \( x_M = R \). Then, for \( i = 1, \ldots, M \), each interval \([x_{i-1}, x_i]\) defines a category \( V_i \) (using the notation in the previous sections). So, all I have to do is specify a total "amount of prior evidence" \( A \) and my prior expectation for each \( \Theta_i \), i.e. my prior probability that a random value from the population falls in \([x_{i-1}, x_i]\), as described in the previous section.

The problem with this approach is that any particular choice of sub-intervals is likely to seem somewhat arbitrary and we might be concerned that different choices could lead to fundamentally different posterior distributions.

4.2 Conglomeration

However, having specified a collection of sub-intervals, we can consider what happens if we combine two intervals (conglomeration). We saw earlier that we can do this by adding the two degrees of freedom and adding the corresponding population proportions.

Suppose we eliminate \( x_1 \) to form a larger interval \([x_0, x_2]\) so that \( \Theta_* = \Theta_1 + \Theta_2 \) is the proportion of the population in \([x_0, x_2]\). By taking \( \alpha_* = \alpha_1 + \alpha_2 \), we ensure consistency in the sense that my new prior mean \( \text{E}[\Theta_*] \) is

\[
\frac{\alpha_*}{A} = \frac{\alpha_1 + \alpha_2}{A} = \frac{\alpha_1}{A} + \frac{\alpha_2}{A}
\]

which is the sum of my former prior means \( \text{E}[\Theta_1] \) and \( \text{E}[\Theta_2] \).

Moreover, since \( k_* = k_1 + k_2 \), my posterior mean \( \text{E}[\Theta_* | \text{data}] \) is

\[
\frac{\alpha_* + k_*}{A + n} = \frac{(\alpha_1 + \alpha_2) + (k_1 + k_2)}{A + n} = \frac{\alpha_1 + k_1}{A + n} + \frac{\alpha_2 + k_2}{A + n}
\]

which is the sum of my former posterior means \( \text{E}[\Theta_1 | \text{data}] \) and \( \text{E}[\Theta_2 | \text{data}] \).

Thus everything is consistent between the original and conglomerated versions since the total amount of evidence \( A + n \) has also not changed.

4.3 Refinement

Alternatively, we could refine the original categorisation by splitting an existing category into two smaller pieces. Suppose we introduce \( x_* \) between \( x_0 \) and \( x_1 \) creating new intervals \([x_0, x_*]\) and \([x_*, x_1]\) with corresponding population proportions \( \Theta_- \) and \( \Theta \) where \( \Theta_1 = \Theta_- + \Theta \).

Provided we choose \( \alpha_- \) and \( \alpha \) so that \( \alpha_1 = \alpha_- + \alpha \), everything remains consistent in exactly the same way as for conglomeration. The prior mean for \( \Theta_1 \) is the sum of the prior means for \( \Theta_- \) and \( \Theta \). The posterior mean for \( \Theta_1 \) is the sum of the posterior means for \( \Theta_- \) and \( \Theta \). The total amount of evidence is unchanged.

4.4 Dirichlet process

If we keep refining the categorisation by subdividing intervals into two smaller pieces as just described, we can obtain an arbitrarily fine partition of the original interval, each sub-interval having its own \( \Theta \) parameter representing the probability of falling in the interval. Prior and posterior means will remain consistent for all the partitions considered provided that at each stage we honour the constraint that the old \( \alpha \) for an interval is the sum of the two new \( \alpha \)s for the sub-intervals.

In the limit, as with integral calculus, we obtain an infinitely fine partition which corresponds to truly non-parametric inference.

When we take this limit, specifying all the prior means for the \( \Theta \) parameters becomes equivalent to specifying a prior mean for the cdf of the population. Basically, we have to write down a prior cdf \( F_0(x) \) and a single number \( A \) (the "amount of prior evidence").

There is much more that could be said but this gives the basic idea of how to develop a truly non-parametric approach to Bayesian inference.