MAS8303 Modern Bayesian Inference
Part 2

M. Farrow
School of Mathematics and Statistics
Newcastle University

Semester 1, 2012-13

Chapter 0

Inference for More Than One Unknown

0.1 More than one unknown

0.1.1 Basic ideas

MAS3301 mostly looked at Bayesian inference in the case where we have a single unknown quantity, usually a parameter. In MAS8303 we will typically look at models with two or more, sometimes many more, unknowns. So, in this lecture, we will look at what happens when we have more than one unknown parameter. The principle is the same when we have more than one parameter. We simply obtain a joint posterior distribution for the parameters. For example, if there are two parameters, we might produce a contour plot of the posterior pdf, as shown in figure 1, or a "3-d" plot, as shown in figure 2. If there are more than two parameters we need to "integrate out" some of the parameters in order to produce graphs like this.

As usual, the basic rule is **posterior** $\propto$ **prior** $\times$ **likelihood.** If necessary, the normalising constant is found by integrating over all parameters. Posterior means, variances, marginal probability density functions, predictive distributions etc. can all be found by suitable integrations. In practice the integrations are often carried out numerically by computer. Apart from being the only practical means in many cases, this removes the pressure to use a convenient conjugate prior.

Sometimes our beliefs might be represented by a model containing several parameters and we might want to answer questions about a number of them. For example, in a medical experiment, we might be interested in the effect of a new treatment on several different outcome measures so we might want to make inferences about the change in the mean for each of these when we move from the old to the new treatment. In frequentist statistics this can give rise to the "multiple testing problem." This problem does not arise for Bayesians. For a Bayesian the inference always consists of the posterior distribution. Once we have calculated the posterior distribution we can calculate whatever summaries we want from it without any logical complications. For example, we could calculate a posterior probability that the mean outcome measure has increased from one treatment to the other for each outcome, or a joint probability that it has increased for every member of some subset of the outcomes or any or all of many other summaries.

0.1.2 The bivariate normal distribution

The normal distribution can be extended to deal with two variables. (In fact, we can extend this to more than two variables).

If Y_1 and Y_2 are two continuous random variables with joint pdf

$$f(\underline{y}) = (2\pi)^{-1}|V|^{-1/2}\exp\left\{-\frac{1}{2}(\underline{y}-\underline{\mu})^T V^{-1}(\underline{y}-\underline{\mu})\right\}$$

for $-\infty < y_1 < \infty$ and $-\infty < y_2 < \infty$ then we say that Y_1 and Y_2 have a bivariate normal

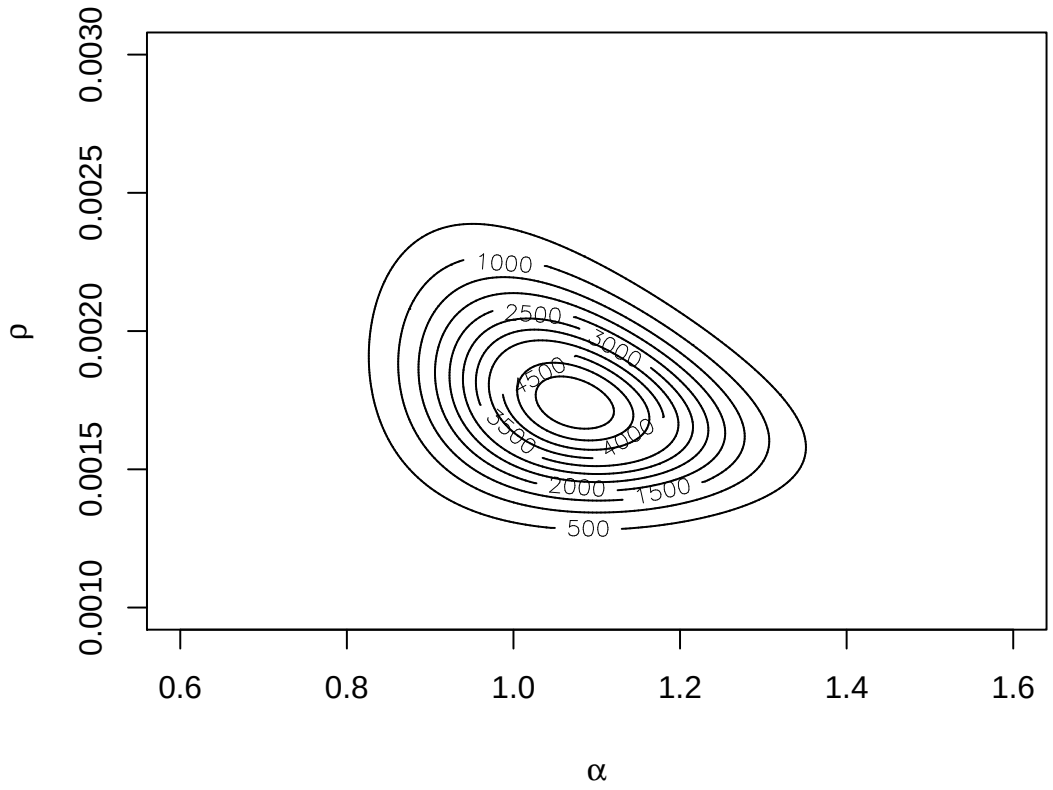


Figure 1: Posterior density of two unknowns: Contour plot

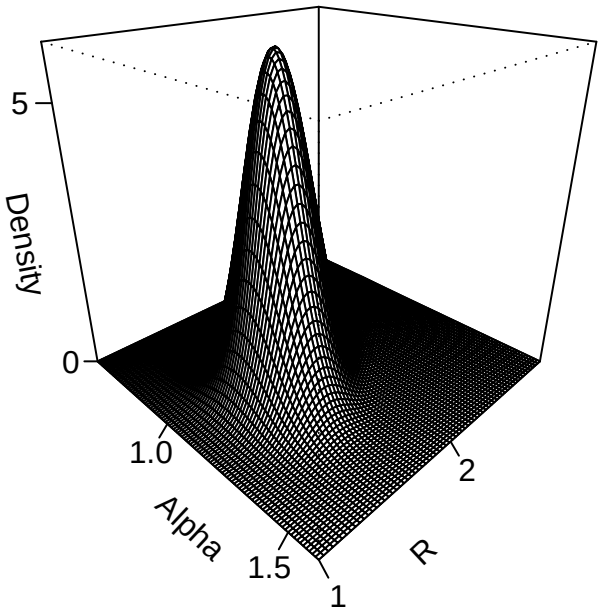


Figure 2: Posterior density of two unknowns: Wireframe plot

distribution with *mean vector* $\underline{\mu} = (\mu_1,\ \mu_2)^T$ and *variance matrix*

$$V = \begin{pmatrix} v_{1,1} & v_{1,2} \\ v_{1,2} & v_{2,2} \end{pmatrix}$$

where μ_1 and μ_2 are the means of Y_1 and Y_2 respectively, $v_{1,1}$ and $v_{2,2}$ are their variances, $v_{1,2}$ is their covariance and $|V|$ is the determinant of V.

If Y_1 and Y_2 are independent then $v_{1,2} = 0$ and, in the case of the bivariate normal distribution, the converse *is* true.

Note that, if X and Y both have normal marginal distributions it does not necessarily follow that their joint distribution is bivariate normal, although, in practice, the joint distribution often is bivariate normal. However, if X and Y both have normal distributions and are independent then their joint distribution is bivariate normal with zero covariance.

If Y_1 and Y_2 have a bivariate normal distribution then $a_1Y_1 + a_2Y_2$ is also normally distributed, where a_1 and a_2 are constants. For example, if $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ and X and Y are independent then $X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

0.1.3 Functions of continuous random variables (Revision)

Theory

As we shall see in the example below, we sometimes need to find the distribution of a random variable which is a function of another random variable. Suppose we have two random variables X and Y where $Y = g(X)$ for some function $g()$. In this section we will only consider the case where $g()$ is a strictly monotonic, i.e. either strictly increasing or strictly decreasing, function.

Suppose first that $g()$ is a strictly increasing function so that if $x_2 > x_1$ then $y_2 = g(x_2) > y_1 = g(x_1)$. In this case the distribution functions $F_X(x)$ and $F_Y(y)$ are related by

$$F_Y(y) = \Pr(Y < y) = \Pr(X < x) = F_X(x).$$

We can find the relationship between the probability density functions , $f_Y(y)$ and $f_X(x)$, by differentiating with respect to y. So

$$f_Y(y) = \frac{d}{dy}F_y(y) = \frac{d}{dy}F_X(x) = \frac{d}{dx}F_X(x) \times \frac{dx}{dy} = f_X(x)\frac{dx}{dy} = f_X(x)\left(\frac{dy}{dx}\right)^{-1}.$$

Similarly, if $g()$ is a strictly decreasing function so that if $x_2 > x_1$ then $y_2 = g(x_2) < y_1 = g(x_1)$,

$$F_Y(y) = \Pr(Y < y) = \Pr(X > x) = 1 - F_X(x)$$

and

$$f_Y(y) = -f_X(x)\frac{dx}{dy}$$

but here, of course, dx/dy is negative.

So, if $g()$ is a strictly monotonic function

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| \text{ where } \left|\frac{dx}{dy}\right| \text{ is the modulus of } \frac{dx}{dy}.$$

A simple way to remember this is to remember that an element of probability $f_X(x)\delta x$ is preserved through the transformation so that (for a strictly increasing function)

$$f_Y(y)\delta y = f_X(x)\delta x.$$

Example

0.1.4 The multivariate normal distribution

Suppose that $\underline{X}$ has a multivariate normal $N_n(\underline{M}, V)$ distribution. This distribution has a *mean vector* $\underline{M} = (m_1, \ldots, m_n)^T$ where m_i is the mean of X_i, and a *covariance matrix* V. The diagonal elements of V are the variances of $X_1, \ldots, X_n$ with the element in row and column i, v_{ii} being the variance of X_i. The covariance of X_i and X_j is v_{ij}, the element in row i and column j. Clearly $v_{ji} = v_{ij}$ and V is symmetric. It is also positive semi-definite.

The pdf is

$$f_X(\underline{x}) = (2\pi)^{-n/2}|V|^{-1/2}\exp\left\{-\frac{1}{2}(\underline{x}-\underline{M})^T V^{-1}(\underline{x}-\underline{M})\right\}.$$

(Here $\underline{x}^T$ denotes the transpose of $\underline{x}$). We often work in terms of the *precision matrix* $P = V^{-1}$. In this case, of course, we replace $(\underline{x}-\underline{M})^T V^{-1}(\underline{x}-\underline{M})$ with $(\underline{x}-\underline{M})^T P(\underline{x}-\underline{M})$.

If $\underline{X}$ has a multivariate normal $N_n(\underline{M},\ V)$ distribution and V is a diagonal matrix, that is if $\text{covar}(X_i, X_j) = 0$ when $i \neq j$, then $X_1, \ldots, X_n$ are independent.

0.1.5 Numerical Methods for More Than One Parameter

It is often necessary to use numerical methods to do the necessary integrations for computing posterior distributions and summaries. Such methods can be used when we have more than one unknown. We will look at this first in the case of two unknown parameters.

If we have two unknown parameters $\theta_1,\ \theta_2$ then we often need to create a two-dimensional grid of values, containing every combination of $\theta_{1,1}, \ldots, \theta_{1,m_1}$ and $\theta_{2,1}, \ldots, \theta_{2,m_2}$, where $\theta_{j,1}, \ldots, \theta_{j,m_j}$ are a set of, usually equally spaced, values of θ_j. We therefore have $m_1 m_2$ points and two step sizes, $\delta\theta_1,\ \delta\theta_2$. Figure 3 shows such a grid diagramatically. Instead of a collection of two-dimensional rectangular columns standing on a one-dimensional line, we now have a collection of three-dimensional rectangular columns standing on a two-dimensional plane. The contours in figure 3 represent the function being integrated. The small circles represent the points at which the function is evaluated. The dashed lines represent the boundaries of the columns. Of course we would really have many more function evaluations placed much more closely together. Notice that some of the function evaluations are in regions where the value of the function is very small. It is inefficient to waste too many function evaluations in this way and some more sophisticated methods avoid doing this.

The approximate integral becomes

$$\int\int h(\theta_1,\theta_2)\ d\theta_1\ d\theta_2 \approx \sum_{j=1}^{m_1}\sum_{k=1}^{m_2} h(\theta_{1,j},\theta_{2,k})\delta\theta_1\delta\theta_2.$$

We can extend this to three or more dimensions but it becomes impractical when the number of dimensions is large. If we use a 100×100 grid in two dimensions this gives 10^4 function evaluations. If we use a $100\times100\times100$ grid in three dimensions this requires 10^6 evaluations and so on. Clearly the number of evaluations becomes prohibitively large quite quickly as the number of dimensions increases. In such cases we would usually use Markov chain Monte Carlo methods which are beyond the scope of this module.

It is sometimes possible to reduce the dimension of the numerical integral by integrating analytically with respect to one unknown.

0.1.6 Example: The Weibull distribution

Model

The *Weibull distribution* is often used as a distribution for *lifetimes*. We might be interested, for example, in the lengths of time that a machine or component runs before it fails, or the survival time of a patient after a serious operation. A number of different families of distributions are used for such lifetime variables. Of course they are all continuous distributions and only give positive probability density to positive values of the lifetime. The Weibull distribution is an important distribution of this type. We can think of it as a generalisation of the exponential distribution. The distribution function of an exponential distribution is $F(t) = 1 - \exp(-\lambda t)$. The distribution function of a Weibull distribution is

$$F(t) = 1 - \exp\left(-\lambda t^\alpha\right) \qquad (t \geq 0) \tag{1}$$

where the extra parameter $\alpha > 0$ is called a *shape parameter*. It is often convenient to write $\lambda = \rho^\alpha$ and then

$$F(t) = 1 - \exp\left(-[\rho t]^\alpha\right) \qquad (t \geq 0) \tag{2}$$

and $\rho > 0$ is a *scale parameter*.

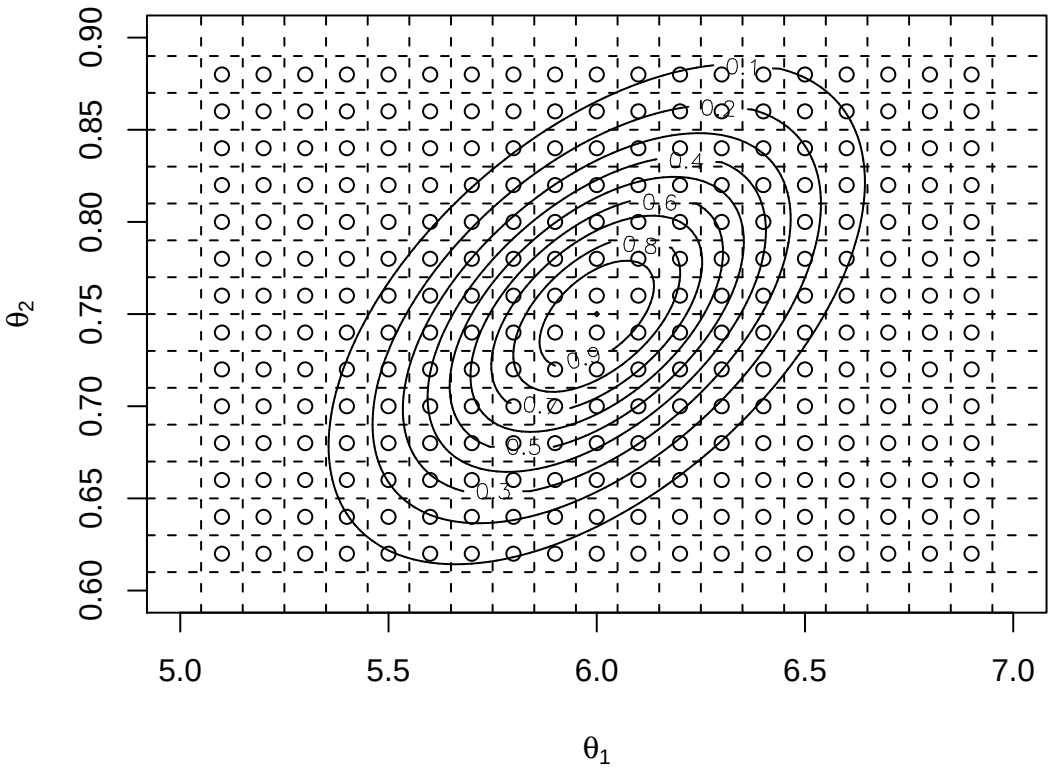


Figure 3: Numerical integration in two dimensions.

Differentiating (2) with respect to t, we obtain the pdf

$$f(t) = \alpha\rho(\rho t)^{\alpha-1}\exp\{-(\rho t)^{\alpha}\} \tag{3}$$

for $0 \le t < \infty$.

If we use $\alpha,\ \lambda$ instead of $\alpha,\ \rho$ as the parameters, as in (1), then the pdf is

$$f(t) = \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha}). \tag{4}$$

Evaluating the posterior distribution

Suppose, for example, that we give α and ρ independent gamma prior distributions so that

$$f^{(0)}_{\alpha,\rho}(\alpha,\rho) \propto \alpha^{a_\alpha-1}e^{-b_\alpha\alpha}\rho^{a_\rho-1}e^{-b_\rho\rho}.$$

Then the posterior pdf is proportional to

$$h_{\alpha,\rho}(\alpha,\rho) = \alpha^{n+a_\alpha-1}\rho^{n\alpha+a_\rho-1}\left(\prod_{i=1}^{n}t_i\right)^{\alpha-1}\exp\left\{-\left[b_\alpha\alpha + b_\rho\rho + \rho^{\alpha}\sum_{i=1}^{n}t_i^{\alpha}\right]\right\}.$$

Figure 1 shows the posterior density of α and ρ when $n = 50,\ a_\alpha = 1,\ b_\alpha = 1,\ a_\rho = 3,\ b_\rho = 1000$ and the data are as given in table 1. Figure 2 shows the same thing as a perspective plot except that, to make the axes more readable, ρ has been replaced with $R = 1000\rho$.

To find, for example, the posterior mean of ρ we evaluate

$$\int_0^\infty\int_0^\infty \rho\ f^{(1)}_{\alpha,\rho}(\alpha,\rho)\ d\alpha\ d\rho = C^{-1}\int_0^\infty\int_0^\infty \rho\ h_{\alpha,\rho}(\alpha,\rho)\ d\alpha\ d\rho.$$

To find a 95 % hpd region for α, ρ we can either choose a value k and evaluate $\int\int f^{(1)}_{\alpha,\rho}(\alpha,\rho)\ d\alpha\ d\rho$ over all points in a grid for which $f^{(1)}_{\alpha,\rho}(\alpha,\rho) > k$ then adjust k and repeat until the value of 0.95

67	313	1391	630	627	573	2093	28	492	482
206	1166	165	1088	496	313	437	815	436	17
32	131	340	939	247	1859	57	132	813	254
950	1615	463	258	2285	672	506	50	637	246
178	431	306	662	33	254	858	187	344	545

Table 1: Data for Weibull example.

is obtained or rank all of the points in our grid in decreasing order of $f^{(1)}_{\alpha,\rho}(\alpha,\rho)$ and cumulatively integrate over them until 0.95 is reached.

To find the marginal pdf for α we evaluate

$$\int_0^\infty f^{(1)}_{\alpha,\rho}(\alpha,\rho)\ d\rho.$$

0.1.7 Transformations

Theory

It has probably become apparent by now that sometimes it may be helpful to use a transformation of the parameters. For example, sometimes a posterior distribution where we need to use numerical integration might have an awkward shape which makes placing a suitable and efficient rectangular grid difficult.

In section 0.1.3 we saw how to change the pdf when we transform a single random variable. Sometimes, of course, we need a more general method for transforming between one set of parameters and another. Let $\underline{\theta}$ and $\underline{\phi}$ be two alternative sets of parameters where there is a 1 - 1 relationship between values of $\underline{\theta}$ and values of $\underline{\phi}$, and therefore each contains the same number of parameters. (There could appear to be more parameters in $\underline{\theta}$ than in $\underline{\phi}$, for example, but, in that case, there would have to be constraints on the values of $\underline{\theta}$ so that there was the same *effective* number of parameters in $\underline{\theta}$ and $\underline{\phi}$). Let $\underline{\theta} = (\theta_1, \ldots, \theta_k)^T$ and $\underline{\phi} = (\phi_1, \ldots, \phi_k)^T$. Suppose also that we can write, for each i,

$$\phi_i = g_i(\theta_1, \ldots, \theta_k)$$

where g is a differentiable function. Then, if the density of $\underline{\theta}$ is $f_{\underline{\theta}}(\underline{\theta})$ and the density of $\underline{\phi}$ is $f_{\underline{\phi}}(\underline{\phi})$,

$$f_{\underline{\theta}}(\underline{\theta}) = f_{\underline{\phi}}(\underline{\phi})|J|$$

where J is the *Jacobian determinant*, often just called "the Jacobian,"

$$\begin{vmatrix} \frac{\partial\phi_1}{\partial\theta_1} & \frac{\partial\phi_1}{\partial\theta_2} & \cdots & \frac{\partial\phi_1}{\partial\theta_k} \\ \frac{\partial\phi_2}{\partial\theta_1} & \frac{\partial\phi_2}{\partial\theta_2} & \cdots & \frac{\partial\phi_2}{\partial\theta_k} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial\phi_k}{\partial\theta_1} & \frac{\partial\phi_k}{\partial\theta_2} & \cdots & \frac{\partial\phi_k}{\partial\theta_k} \end{vmatrix}$$

and $|J|$ is its modulus.

For example, we could transform the $(0,\infty)$ ranges of the parameters α, ρ of a Weibull distribution to $(0,1)$ by using

$$\beta = \frac{\alpha}{\alpha+1}, \qquad \gamma = \frac{\rho}{\rho+1}.$$

The Jacobian is

$$J = \begin{vmatrix} \frac{\partial\beta}{\partial\alpha} & \frac{\partial\beta}{\partial\rho} \\ \frac{\partial\gamma}{\partial\alpha} & \frac{\partial\gamma}{\partial\rho} \end{vmatrix} = (\alpha+1)^{-2}(\rho+1)^{-2}.$$

Suppose that the joint posterior density of α and ρ is proportional to $h_{\alpha,\rho}(\alpha,\rho)$. So we define

$$h_{\beta,\gamma}(\beta,\gamma) = (\alpha+1)^2(\rho+1)^2 h_{\alpha,\rho}(\alpha,\rho),$$

where

$$\alpha = \frac{\beta}{1-\beta}, \qquad \rho = \frac{\gamma}{1-\gamma}$$

so

$$h_{\beta,\gamma}(\beta,\gamma) = (1-\beta)^{-2}(1-\rho)^{-2} h_{\alpha,\rho}\left(\frac{\beta}{1-\beta}, \frac{\gamma}{1-\gamma}\right).$$

Then let

$$C = \int_0^1 \int_0^1 h_{\beta,\gamma}(\beta,\gamma)\ d\beta\ d\gamma.$$

The posterior mean of ρ is then

$$C^{-1} \int_0^1 \int_0^1 \frac{\gamma}{1-\gamma} h_{\beta,\gamma}(\beta,\gamma)\ d\beta\ d\gamma.$$

A hpd region for α, ρ can then be found by integrating $C^{-1}h_{\beta,\gamma}(\beta,\gamma)$ with respect to β, γ over the points with the greatest values of

$$h_{\alpha,\rho}\left(\frac{\beta}{1-\beta}, \frac{\gamma}{1-\gamma}\right) = h_{\alpha,\rho}(\alpha,\rho).$$

Example: A clinical trial

The Anturane Reinfarction Trial Research Group (1980) reported a clinical trial on the use of the drug sulfinpyrazone in patients who had suffered myocardial infarctions ("heart attacks"). The idea was to see whether the drug had an effect on the number dying. Patients in one group were given the drug while patients in another group were given a "placebo," that is an inactive substitute. The following table gives the number of all "analysable" deaths up to 24 months after the myocardial infarction and the total number of eligible patients who were not withdrawn and did not suffer a "non-analysable" death during the study.

	Deaths	Total
Group 1 (Sulfinpyrazone)	44	560
Group 2 (Placebo)	62	540

We can represent this situation by saying that there are two groups, containing n_1 and n_2 patients, and two parameters, $\theta_1,\ \theta_2$, such that, given these parameters, the distribution of the number of deaths X_j in Group j is binomial$(n_j,\ \theta_j)$.

Now we could give θ_j a beta prior distribution but it seems reasonable that our prior beliefs would be such that θ_1 and θ_2 would not be independent. There are various ways in which we could represent this. One of these is as follows. We transform from the $(0,1)$ scale of $\theta_1,\ \theta_2$ to a $(-\infty,\infty)$ scale and then give the new parameters, $\eta_1,\ \eta_2$, a bivariate normal distribution (see section 0.1.2). We can use a transformation where $\theta_j = F(\eta_j)$ and $F(x)$ is the distribution function of a continuous distribution on $(-\infty,\ \infty)$, usually one which is symmetric about $x = 0$. One possibility is to use the standard normal distribution function $\Phi(x)$ so that $\theta_j = \Phi(\eta_j)$. We write $\eta_j = \Phi^{-1}(\theta_j)$ where this function, $\Phi^{-1}(x)$, the inverse of the standard normal distribution function, is sometimes called the *probit* function. If we use this transformation then it is easily seen that

$$f_\theta(\theta_1,\theta_2) = f_\eta(\eta_1,\eta_2)/|J|,$$

where $f_\theta(\theta_1,\theta_2)$ is the joint density of $\theta_1,\ \theta_2$, $f_\eta(\eta_1,\ \eta_2)$ is the joint density of $\eta_1,\ \eta_2$ and

$$|J| = \left| \begin{array}{cc} \frac{\partial\theta_1}{\partial\eta_1} & \frac{\partial\theta_1}{\partial\eta_2} \\ \frac{\partial\theta_2}{\partial\eta_1} & \frac{\partial\theta_2}{\partial\eta_2} \end{array} \right| = \phi(\eta_1)\phi(\eta_2),$$

where $\phi(x)$ is the standard normal pdf.

Suppose that, from past experience, we can give a 95% symmetric prior interval for θ_2 (placebo) as $0.05 < \theta_2 < 0.20$. (This is actually quite a wide interval considering that there may be a lot of past experience of such patients). This converts to a 95% interval of $-1.645 < \eta_2 < -0.842$. For example, in R, we can use

```
> qnorm(0.025,0,1)
[1] -1.959964
```

If we give η_2 a normal prior distribution then we require the mean to be $\mu_2 = ([-1.645] + [-0.842])/2 \approx -1.24$ and the standard deviation to be $\sigma_2 = ([-0.842]-[-1.645])/(2\times 1.96) \approx 0.21$, since a symmetric 95% normal interval is the mean plus or minus 1.96 standard deviations. Let us use the same mean for a normal prior distribution for η_1 (sulfinpyrazone) so that we have equal prior probabilities for an increase and a decrease in death rate when the treatment is given. However it seems reasonable that we would be less certain of the death rate given the treatment so we increase the prior standard deviation to $\sigma_1 = 2\sigma_2 = 0.42$. This implies a 95% interval $-2.06 < \eta_1 < -0.42$ which, in turn, implies $0.02 < \theta_1 < 0.34$. (This is a wide interval so we are really not supplying much prior information).

We also need to choose a covariance or correlation between η_1 and η_2. At this point we will not discuss in detail how to do this except to say that, if we choose the correlation to be r, then the conditional variance of one of $\eta_1,\ \eta_2$ given the other will be $100r^2\%$ of the marginal variance. For example, if we choose $r = 0.7$, then the variance of one is roughly halved by learning the value of the other. Suppose that we choose this value. Then the covariance between η_1 and η_2 is $0.7 \times 0.21 \times 0.42 = 0.0617$.

In evaluating the joint prior density of $\eta_1,\ \eta_2$, we can make use of the fact, which is easily confirmed, that, if $\delta_j = (\eta_j - \mu_j)/\sigma_j$ and $r = \text{covar}(\eta_1, \eta_2)/(\sigma_1\sigma_2)$, then the joint density is proportional to

$$\exp\left\{-\frac{1}{2(1-r^2)}(\delta_1^2 + \delta_2^2 - 2r\delta_1\delta_2)\right\}.$$

Figure 4 shows a R function to evaluate the posterior density. Figure 5 shows the resulting posterior density. The dashed line is the line $\theta_1 = \theta_2$. We see that most of the probability lies on the side where $\theta_2 > \theta_1$ which suggests that the death rate is probably greater with the placebo than with sulfinpyrazone, which, of course, suggests that sulfinpyrazone has a beneficial effect.

To investigate further what the posterior tells us about the effect of sulfinpyrazone, we can calculate the posterior probability that $\theta_1 < \theta_2$. This is done by integrating the joint posterior density over the region where $\theta_1 < \theta_2$. This calculation is included in the function shown in figure 4. The calculated probability is 0.972. We can also find the posterior density of the *relative risk*, θ_1/θ_2, or the *log relative risk*, $\log(\theta_1/\theta_2)$. Let γ be the log relative risk. We can modify the function in figure 4 so that it uses a grid of γ and θ_2 values, evaluates the joint posterior density of γ and θ_2 and then integrates out θ_2. Of course we need to transform between $\theta_1,\ \theta_2$ and $\gamma,\ \theta_2$ where the densities are related by

$$f_{\theta_1,\theta_2}(\theta_1, \theta_2) = f_{\gamma,\theta_2}(\gamma, \theta_2)|J|$$

and $J = \theta_1^{-1}$ is the appropriate Jacobian. Figure 6 shows the prior and posterior densities of the log relative risk, γ. Values of γ less than zero correspond to a smaller death rate with sulfinpyrazone than with the placebo. Notice that the prior density is not quite symmetric about zero. It is symmetric on the η scale but not on the γ scale. The prior median is zero, however.

There are other methods available to deal with problems of this sort, some involving approximations and fairly simple calculations.

```
function(theta1,theta2,n,x,prior)
{# Evaluates posterior density for probit example.
 # prior is mean1, mean2, sd1, sd2, correlation
 n1<-length(theta1)
 n2<-length(theta2)
 step1<-theta1[2]-theta1[1]
 step2<-theta2[2]-theta2[1]
 theta1<-matrix(theta1,nrow=n1,ncol=n2)
 theta2<-matrix(theta2,nrow=n1,ncol=n2,byrow=T)
 eta1<-qnorm(theta1,0,1)
 eta2<-qnorm(theta2,0,1)
 delta1<-(eta1-prior[1])/prior[3]
 delta2<-(eta2-prior[2])/prior[4]
 r<-prior[5]
 d<-1-r^2
 logprior<- -(delta1^2 + delta2^2 - 2*r*delta1*delta2)/(2*d)
 J<-dnorm(eta1,0,1)*dnorm(eta2,0,1)
 logprior<-logprior-log(J)
 loglik<-x[1]*log(theta1)+(n[1]-x[1])*log(1-theta1)+x[2]*log(theta2)+(n[2]-x[2])*log(1-theta2)
 logpos<-logprior+loglik
 logpos<-logpos-max(logpos)
 posterior<-exp(logpos)
 int<-sum(posterior)*step1*step2
 posterior<-posterior/int
 prob<-sum(posterior*(theta1<theta2))*step1*step2
 ans<-list(density=posterior,prob=prob)
 ans
 }
```

Figure 4: R function for probit example (0.1.7).

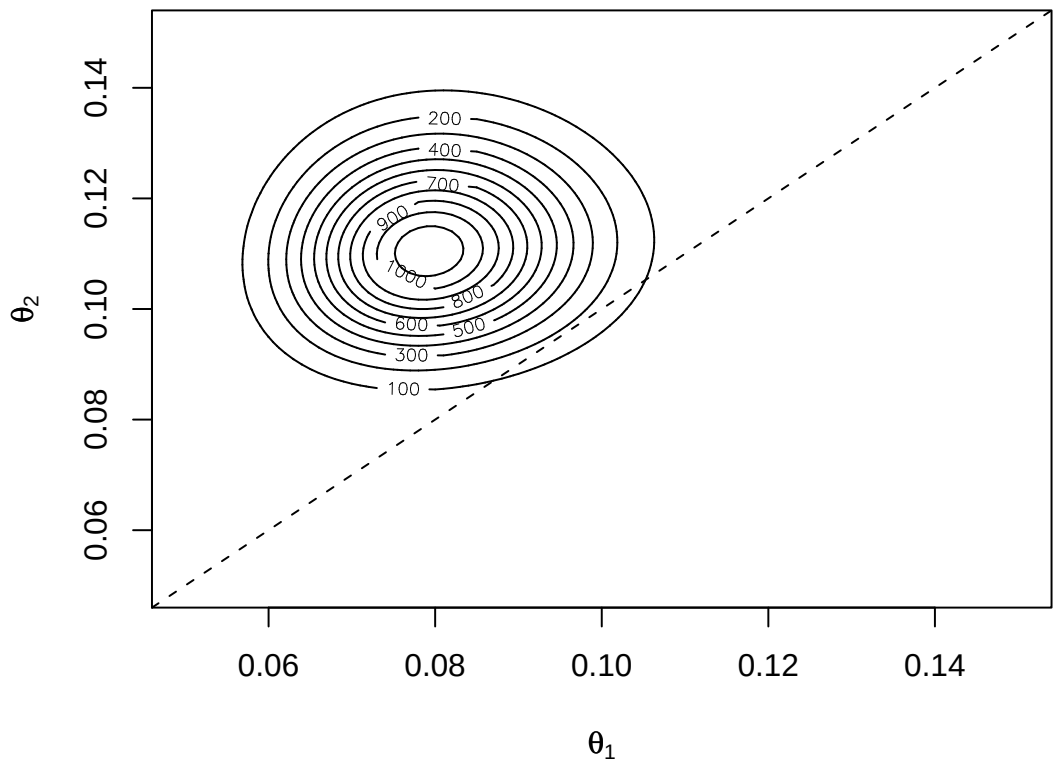


Figure 5: Posterior density of θ_1 and θ_2 in probit example (0.1.7). The dashed line is $\theta_1 = \theta_2$.

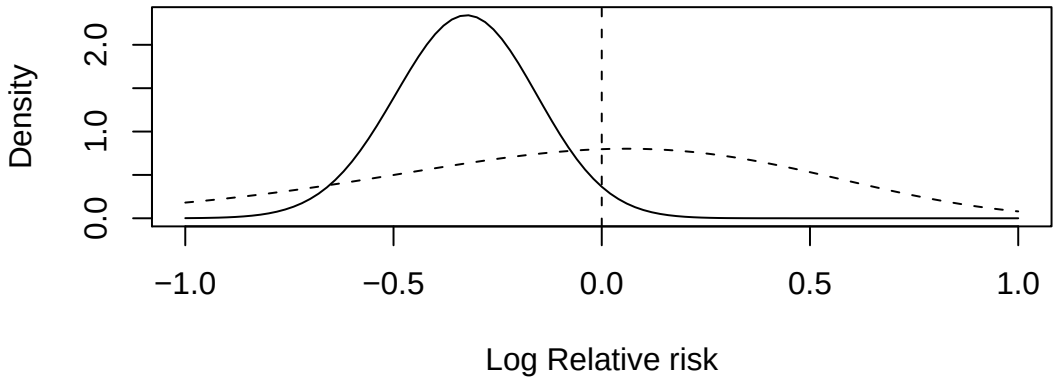


Figure 6: Posterior density (solid) and prior density (dashes) of log relative risk in probit example (0.1.7).

0.2 The Dirichlet distribution and multinomial observations

0.2.1 The Dirichlet distribution

The Dirichlet distribution is a distribution for a set of quantities $\theta_1, \ldots, \theta_m$ 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

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

where $A = \sum_{i=1}^m a_i$ and $a_1, \ldots, a_m$ are parameters with $a_i > 0$ for $i = 1, \ldots, m$. We write $\mathrm{D}_m(a_1, \ldots, a_m)$ for this distribution.

Clearly, if $m = 2$, we obtain a $\mathrm{Beta}(a_1, a_2)$ 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

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

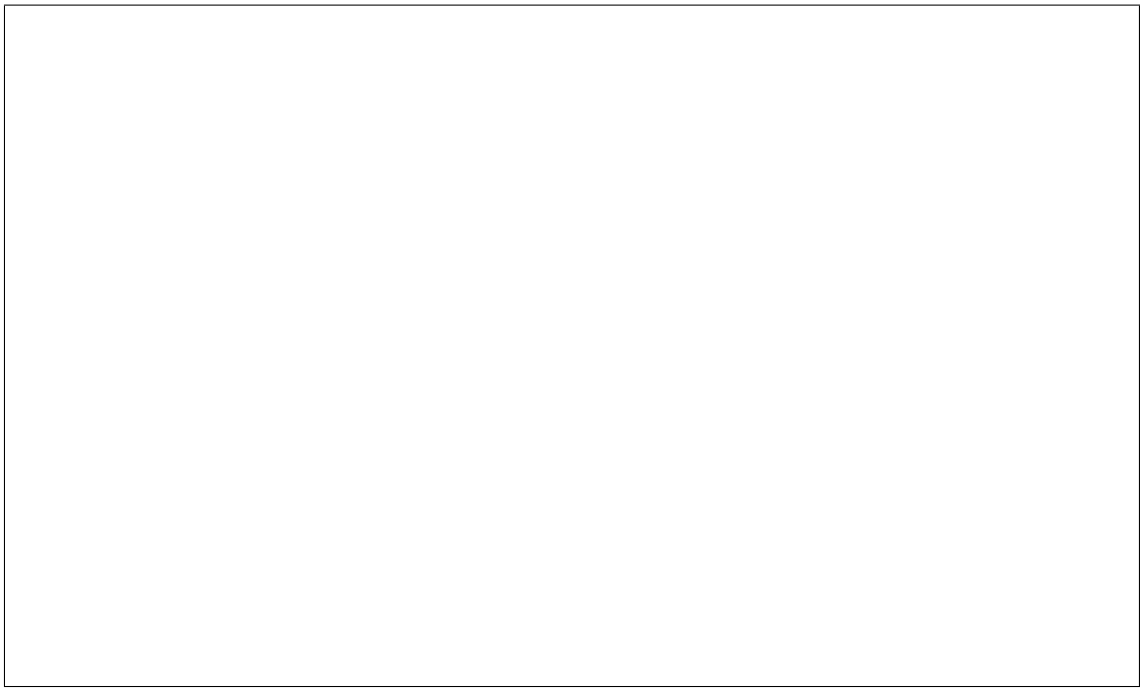
and the covariance of θ_j and θ_k, where $j \neq k$, is

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

Also the marginal distribution of θ_j is $\mathrm{Beta}(a_j,\ A - a_j)$.

Note that the space of the parameters $\theta_1, \ldots, \theta_m$ 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)



Proof (variance)

Proof (covariance)

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 $\mathrm{D}_d(a_1,\ a_2,\ a_3,\ \tilde{a}_3)$ where $\tilde{a}_3 = A-a_1-a_2-a_3$.

0.2.2 Multinomial observations

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 Bin(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 $a_1+x_1, a_2+x_2, \ldots a_m+x_m$.

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 $\mathrm{D}_6(5,\ 3,\ 5,\ 1,\ 2,\ 4)$ 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		