Chapter 5

More Bayesian Inference

We use the generic $p(\cdot)$ notation for densities, if there is no danger of confusion.

5.1 Likelihoods and sufficient statistics

Let us consider n (conditionally) independent Bernoulli trials $Y_1, \dots, Y_n$ with success probability θ. That is, the RVs Y_i are independent and Y_i takes on the value 1 with probability θ (success in the i'th Bernoulli experiment) and otherwise is zero (failure in the i'th Bernoulli experiment). Having observed the values $y_1, \dots, y_n$, the likelihood corresponding to $y = (y_1, \dots, y_n)$ is given by

$$\begin{aligned} p(y \mid \theta) &= \prod_{i=1}^{n} p(y_i \mid \theta) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \\ &= \theta^s (1-\theta)^{n-s}, \qquad 0 < \theta < 1, \end{aligned} \tag{5.1}$$

where

$$s = t(y) = \sum_{i=1}^{n} y_i$$

is the observed number of successes. Here the likelihood depends on the data y only through the value of $t(y)$, which is said to be a **sufficient statistic**. Since

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta) = \theta^{t(y)} (1-\theta)^{n-t(y)}\, p(\theta),$$

the posterior depends on the data only through the value of $t(y)$.

In a more general situation, a statistic $t(Y)$ is called sufficient, if the likelihood can be factored as

$$p(y \mid \theta) = g(t(y), \theta)\, h(y)$$

for some functions g and h. Then (as a function of θ)

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta) \propto g(t(y), \theta)\, p(\theta)$$

and therefore the posterior depends on the data only through the value $t(y)$ of the sufficient statistic.

In other words, we might as well throw away the original data as soon as we have calculated the value of the sufficient statistic. (Do not try this at home. You might later want to consider other likelihoods for your data!) Sufficient statistics are very convenient, but not all likelihoods admit a sufficient statistic of a fixed dimension, when the sample size is allowed to vary. Such sufficient statistics exist only in what are known as exponential families, see, e.g., the text of Schervish [6, Ch. 2] for a discussion.

In the Bernoulli trial example, the random variable S corresponding to the sufficient statistic

$$S = t(Y) = \sum_{i=1}^{n} Y_i$$

has the binomial distribution $\mathrm{Bin}(n, \theta)$ with sample size n and success probability θ. I.e., if we observe only the number of success s (but not the order in which the successes and failures happened), then the likelihood is given by

$$p(s \mid \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}, \qquad 0 < \theta < 1. \tag{5.2}$$

The two functions (5.1) and (5.2) describe the same experiment, and are proportional to each other (as functions of θ). The difference stems from the fact that there are exactly $\binom{n}{s}$ equally probable sequences $y_1, \ldots, y_n$, which sum to a given value of s, where $0 \le s \le n$. Since the two functions are proportional to each other, we will get the same posterior with either of them if we use the same prior. Therefore it does not matter which of the expressions (5.1) and (5.2) we use as the likelihood for a binomial experiment.

Observations.

- When calculating the posterior, you can always leave out from the likelihood such factors, which depend only on the data but not on the parameter. Doing that does not affect the posterior.
- If your model admits a convenient sufficient statistic, you do not need to work out the distribution of the sufficient statistic in order to write down the likelihood. You can always use the likelihood of the underlying repeated experiment, even if the original data has been lost and only the sufficient statistic has been recorded.
- However, if you do know the density of the sufficient statistic (conditionally on the parameter), you can use that as the likelihood. (This is tricky; consult, e.g., Schervish [6, Ch. 2] for a proof.)

We can generalize the Bernoulli experiment (or binomial experiment) to the case, where there are $k \ge 2$ possible outcomes instead of two possible outcomes. Consider an i.i.d. sample $Y_1, \ldots, Y_n$ from the discrete distribution with k different values $1, \ldots, k$ with respective probabilities $\theta_1, \ldots, \theta_k$, where $0 < \theta_j < 1$ and $\sum \theta_j = 1$. (Because of the sum constraint, there are actually only $k-1$ free parameters.) The likelihood corresponding to the data $y = (y_1, \ldots, y_n)$ is given by

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \theta_j^{1(y_i = j)} = \theta_1^{n_1} \, \theta_2^{n_2} \cdots \theta_k^{n_k}, \tag{5.3}$$

where n_j is the number of y_is which take on the value j. This is the **multinomial likelihood**. Clearly the frequencies $n_1, \ldots, n_k$ form a sufficient statistic. Notice that $\sum_j n_j = n$.

In this case it is possible to work out the distribution of the sufficient statistic, i.e., the random frequency vector $N = (N_1, \ldots, N_k)$, where

$$N_j = \#\{i = 1, \ldots, n : Y_i = j\}, \qquad j = 1, \ldots, k.$$

Using combinatorial arguments it can be easily proven that

$$\begin{aligned} &P(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k \mid \theta_1, \theta_2, \ldots, \theta_k) \\ &\qquad = \binom{n}{n_1, n_2, \cdots, n_k} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_k^{n_k}, \end{aligned} \tag{5.4}$$

when the integers $0 \le n_1, \ldots, n_k \le n$ and $\sum_j n_j = n$. Here

$$\binom{n}{n_1, n_2, \cdots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!} \tag{5.5}$$

is called a **multinomial coefficient**. The multivariate discrete distribution with pmf (5.4) is called the **multinomial distribution** with sample size parameter n and probability vector parameter $(\theta_1, \ldots, \theta_k)$. The binomial distribution is a special case of the multinomial distribution: if $S \sim \mathrm{Bin}(n, p)$, then the vector $(S, n - S)$ has the multinomial distribution with parameters n and $(p, 1 - p)$.

Notice that we can use the simple expression (5.3) for the likelihood of a multinomial observation even when we know very well that the pmf of the random vector $(N_1, \ldots, N_k)$ involves the multinomial coefficient.

5.2 Conjugate analysis

Some likelihoods have the property that if the prior is selected from a certain family of distributions, then the posterior also belongs to the same family. Such a family is called closed under sampling or a conjugate family (for the likelihood under consideration). A trivial and useless example of a conjugate family is provided by the set of all distributions. The useful conjugate families can be described by a finite number of hyperparameters, i.e., they are of the form

$$\{\theta \mapsto f(\theta \mid \phi) : \phi \in S\}, \tag{5.6}$$

where S a set in an Euclidean space, and $\theta \mapsto f(\theta \mid \phi)$ is a density for each value of the hyperparameter vector $\phi \in S$. If the likelihood $p(y \mid \theta)$ admits this conjugate family, and if the prior $p(\theta)$ is $f(\theta \mid \phi_0)$ with a known value ϕ_0, then the posterior is of the form

$$\theta \mapsto p(\theta \mid y) = f(\theta \mid \phi_1),$$

where $\phi_1 \in S$. In order to find the posterior, we only need to find the value of the updated hyperparameter vector $\phi_1 = \phi_1(y)$.

If the densities $f(\theta \mid \phi)$ of the conjugate family have an easily understood form, then Bayesian inference is simple, provided we can approximate our prior knowledge by some member $f(\theta \mid \phi_0)$ of the conjugate family and provided we know how to calculate the updated hyperparameters $\phi_1(y)$. However, nice conjugate families of the form (5.6) are possible only when the likelihood belongs to the exponential family, see, e.g., Schervish [6, Ch. 2].

The prior knowledge of the subject matter expert on θ is, unfortunately, usually rather vague. Transforming the subject matter expert's prior knowledge into a prior distribution is called **prior elicitation**. See, e.g., [3] for examples of how this could be achieved for certain important statistical models. Supposing we are dealing with a scalar parameter, the expert might only have a feeling for the order of magnitude of the parameter, or might be able to say, which values would be surprisingly small or surprisingly large for the parameter. One approach for constructing the prior would then be to select from the family (5.6) some prior, which satisfies those kind of prior summaries.

As an example of conjugate analysis, consider the binomial likelihood (5.1) corresponding to sample size n and success probability θ. Recall that the beta density with (hyper)parameters $a, b > 0$ is given by

$$\mathrm{Be}(\theta \mid a, b) = \frac{1}{B(a,b)}\, \theta^{a-1}(1-\theta)^{b-1}, \qquad 0 < \theta < 1.$$

Suppose that the parameter θ has the beta prior $\mathrm{Be}(a,b)$ with known hyperparameters a and b. Then

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)\, p(\theta) \\ &\propto \theta^{s}\,(1-\theta)^{n-s}\,\theta^{a-1}\,(1-\theta)^{b-1} \\ &\propto \mathrm{Be}(\theta \mid a+s, b+n-s), \qquad 0 < \theta < 1. \end{aligned}$$

Therefore we claim that the posterior is $\mathrm{Be}(a+s, b+n-s)$, where s is the number of successes (and $n-s$ is the number of failures). Notice the following points.

- We developed the posterior density, as a function of the parameter θ, dropping any constants (i.e., factors not involving θ).

- It is important to keep in mind, which is the variable we are interested in and what are the other variables, whose functions we treat as constants. The variable of interest is the one whose posterior distribution we want to calculate.

- We finished the calculation by recognizing that the posterior has a familiar functional form. In the present example we obtained a beta density except that it did not have the right normalizing constant. However, the only probability density on $0 < \theta < 1$ having the derived functional form is the beta density $\mathrm{Be}(\theta \mid a+s, b+n-s)$, and therefore the posterior distribution is this beta distribution.

- In more detail: from our calculations, we know that the posterior has the unnormalized density $\theta^{a+s-1}(1-\theta)^{b+n-s-1}$ on $0 < \theta < 1$. Since we know that the posterior density is a density on $(0,1)$, we can find the normalizing constant by integration:

$$p(\theta \mid y) = \frac{1}{c(y)}\, \theta^{a+s-1}(1-\theta)^{b+n-s-1}, \qquad 0 < \theta < 1,$$

where

$$c(y) = \int_0^1 \theta^{a+s-1}(1-\theta)^{b+n-s-1}\,\mathrm{d}\theta = B(a+s, b+n-s),$$

where the last step is immediate, since the integral is the normalizing constant of the beta density $\mathrm{Be}(\theta \mid a_1, b_1)$, where $a_1 = a+s$ and $b_1 = b+n-s$. Therefore

$$p(\theta \mid y) = \mathrm{Be}(\theta \mid a+s, b+n-s).$$

- As soon as we have recognized the functional form of the posterior, we have recognized the posterior distribution.

5.3 More examples of conjugate analysis

5.3.1 Poisson sampling model and gamma prior

Suppose that

$$Y_i \mid \theta \overset{\text{i.i.d.}}{\sim} \mathrm{Poi}(\theta), \qquad i = 1, \dots, n,$$

which is shorthand notation for the statement that the RVs $Y_i, i = 1, \dots, n$ are independently Poisson distributed with parameter θ. Then

$$p(y_i \mid \theta) = \frac{1}{y_i!}\,\theta^{y_i}\,\mathrm{e}^{-\theta}, \qquad y_i = 0, 1, 2, \dots$$

and the likelihood is given by

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) \propto \theta^{\sum_1^n y_i}\,\mathrm{e}^{-n\theta}.$$

The likelihood has the functional form of a gamma density. If the prior for θ is the gamma distribution $\mathrm{Gam}(a, b)$ with known hyperparameters $a, b > 0$, i.e., if

$$p(\theta) = \frac{b^a}{\Gamma(a)}\,\theta^{a-1}\mathrm{e}^{-b\theta}, \qquad \theta > 0,$$

then

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)\,p(\theta) \\ &\propto \theta^{\sum_1^n y_i}\mathrm{e}^{-n\theta}\theta^{a-1}\mathrm{e}^{-b\theta} \\ &\propto \theta^{a+\sum_1^n y_i - 1}\,\mathrm{e}^{-\theta(b+n)}, \qquad \theta > 0 \end{aligned}$$

and from this we recognize that the posterior is the gamma distribution

$$\mathrm{Gam}(a + \sum_1^n y_i, b+n).$$

5.3.2 Exponential sampling model and gamma prior

Suppose that

$$\begin{aligned} Y_i \mid \theta &\overset{\text{i.i.d.}}{\sim} \operatorname{Exp}(\theta), \qquad i = 1, \ldots, n \\ \Theta &\sim \operatorname{Gam}(a, b), \end{aligned}$$

where $a, b > 0$ are known constants. Then

$$p(y_i \mid \theta) = \theta \, \mathrm{e}^{-\theta y_i}, \qquad y_i > 0,$$

and the likelihood is

$$p(y \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) = \theta^n \, \exp(-\theta \sum_{i=1}^{n} y_i).$$

We obtain $\operatorname{Gam}(a + n, b + \sum y_i)$ as the posterior.

5.4 Conjugate analysis for univariate normal observations

5.4.1 Known variance but unknown mean

Suppose that we have one normally distributed observation $Y \sim N(\theta, \tau^2)$, where the mean θ is unknown but the variance τ^2 is a known value. Then

$$p(y \mid \theta) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y-\theta)^2}{\tau^2}\right).$$

Suppose that the prior is $N(\mu_0, \sigma_0^2)$ with known constants μ_0 and σ_0^2. Then the posterior is

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta) \, p(\theta) \\ &\propto \exp\left(-\frac{1}{2\tau^2}(y-\theta)^2 - \frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right) = \exp\left(-\frac{1}{2}\, q(\theta)\right), \end{aligned}$$

where

$$q(\theta) = \frac{1}{\tau^2}(y-\theta)^2 + \frac{1}{\sigma_0^2}(\theta-\mu_0)^2$$

is a second degree polynomial in θ, and the coefficient of θ^2 in $q(\theta)$ is positive. Therefore the posterior is a certain normal distribution. However, we need to calculate its mean μ_1 and variance σ_1^2. This we achieve by *completing the square* in the quadratic polynomial $q(\theta)$. However, we need only to keep track of the first and second degree terms.

Developing the density $N(\theta \mid \mu_1, \sigma_1^2)$ as a function of θ, we obtain

$$\begin{aligned} N(\theta \mid \mu_1, \sigma_1^2) &= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\theta-\mu_1)^2}{\sigma_1^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_1^2}\,\theta^2 - 2\,\frac{\mu_1}{\sigma_1^2}\,\theta\right)\right) \end{aligned}$$

Next, we equate the coefficients of θ^2 and θ, firstly, in $q(\theta)$ and, secondly, in the previous formula to find out that we have

$$p(\theta \mid y) = N(\theta \mid \mu_1, \sigma_1^2),$$

where

$$\frac{1}{\sigma_1^2} = \frac{1}{\tau^2} + \frac{1}{\sigma_0^2}, \qquad \frac{\mu_1}{\sigma_1^2} = \frac{y}{\tau^2} + \frac{\mu_0}{\sigma_0^2}, \tag{5.7}$$

from which we can solve first σ_1^2 and then μ_1.

In Bayesian inference it is often convenient to parametrize the normal distribution by its mean and precision, where precision is defined as the reciprocal of the variance. We have just shown that the posterior precision equals the prior precision plus the datum precision.

If we have n independent observations $Y_i \sim N(\theta, \tau^2)$ with a known variance, then it is a simple matter to show that

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

is a sufficient statistic. In this case we know the distribution of the corresponding RV $\bar{Y}$ conditionally on θ,

$$[\bar{Y} \mid \theta] \sim N(\theta, \frac{\tau^2}{n}),$$

From these two facts we get immediately the posterior distribution from (5.7), when the prior is again $N(\mu_0, \sigma_0^2)$. (Alternatively, we may simply multiply the likelihood with the prior density, and examine the resulting expression.)

5.4.2 Known mean but unknown precision

Suppose that the RVs Y_i are independently normally distributed,

$$Y_i \mid \theta \overset{\text{i.i.d.}}{\sim} N(\mu, \frac{1}{\theta}), \qquad i = 1, \ldots, n$$

where the mean μ is known but the variance $1/\theta$ is unknown. Notice that we parametrize the sampling distribution using the precision θ instead of the variance $1/\theta$. Then

$$p(y_i \mid \theta) = \frac{\sqrt{\theta}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\theta(y_i - \mu)^2\right),$$

and the likelihood is

$$p(y \mid \theta) \propto \theta^{n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \theta\right).$$

If the prior is $\text{Gam}(a, b)$, then the posterior is evidently

$$\text{Gam}(a + \frac{n}{2}, b + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2).$$

The previous result can be expressed also in terms of the variance $\phi = 1/\theta$. The variance has what is known as the inverse gamma distribution with density

$$\mathrm{Gam}(\frac{1}{\phi} \mid a_1, b_1)\, \frac{1}{\phi^2}, \qquad \phi > 0,$$

where a_1 and b_1 are the just obtained updated parameters, as can be established by the change of variable $\phi = 1/\theta$ in the posterior density. The inverse gamma distribution is also called the scaled inverse chi-square distribution, using a certain other convention for the parametrization.

5.4.3 Both the mean and the precision are unknown

Suppose that the RVs Y_i are independently normally distributed with unknown mean ϕ and unknown precision τ,

$$[Y_i \mid \phi, \tau] \overset{\text{i.i.d.}}{\sim} N(\phi, \frac{1}{\tau}), \qquad i = 1, \ldots, n.$$

In this case the likelihood for $\theta = (\phi, \tau)$ is conjugate for the prior of the form

$$p(\phi, \tau \mid a_0, b_0, \mu_0, n_0) = \mathrm{Gam}(\tau \mid a_0, b_0)\, N(\phi \mid \mu_0, \frac{1}{n_0 \tau}).$$

Notice that the precision and the mean are dependent in this prior. This kind of a dependent prior may be natural in some problems but less natural in some other problems.

Often the interest centers on the mean ϕ while the precision τ is regarded as a nuisance parameter. The marginal posterior of ϕ (i.e., the marginal distribution of ϕ in the joint posterior) is obtained from the joint posterior by integrating out the nuisance parameter,

$$p(\phi \mid y) = \int p(\phi, \tau \mid y)\, \mathrm{d}\tau.$$

In the present case, this integral can be solved analytically, and the marginal posterior of ϕ can be shown to be a t-distribution.

5.5 Conjugate analysis for the multivariate normal sampling model

When dealing with the multivariate instead of the univariate normal distribution, it is even more convenient to parametrize the normal distribution using the precision matrix, which is defined as the inverse of the covariance matrix, which we assume to be non-singular. Like the covariance matrix, also the precision matrix is a symmetric and positive definite matrix.

The density of the multivariate normal $N_d(\mu, Q^{-1})$ with mean μ and precision matrix Q (i.e., of $N_d(\mu, \Sigma)$, where the covariance matrix $\Sigma = Q^{-1}$) is given by

$$N_d(x \mid \mu, Q^{-1}) = (2\pi)^{-d/2} (\det Q)^{1/2} \exp\left(-\frac{1}{2}(x-\mu)^T Q (x-\mu)\right)$$

where d is the dimensionality of x.

5.5.1 Unknown mean vector but known precision matrix

Expanding the quadratic form inside the exponential function in $N_d(x \mid \mu, Q^{-1})$, we get

$$(x-\mu)^T Q(x-\mu) = x^T Qx - x^T Q\mu - \mu^T Qx + \mu^T Q\mu$$

Now, the precision matrix Q is symmetric, and a scalar equal its transpose, so

$$\mu^T Qx = (\mu^T Qx)^T = x^T Q^T \mu = x^T Q\mu.$$

That is

$$(x-\mu)^T Q(x-\mu) = x^T Qx - 2x^T Q\mu + \mu^T Q\mu$$

(for any symmetric matrix Q), which should be compared with the familiar formula $(a-b)^2 = a^2 - 2\,ab + b^2$ valid for scalars a and b.

Therefore, as a function of x,

$$N_d(x \mid \mu, Q^{-1}) \propto \exp\left(-\frac{1}{2}(x^T Qx - 2x^T Q\mu)\right). \tag{5.8}$$

Suppose that we have a single multivariate observation $Y \sim N(\theta, R^{-1})$, where the prior precision matrix R is known and suppose that the prior for the parameter vector θ is the normal distribution $N(\mu_0, Q_0^{-1})$ with known hyperparameters μ_0 and Q_0. Then

$$p(y \mid \theta) \propto \exp\left(-\frac{1}{2}(y-\theta)^T R(y-\theta)\right).$$

The prior is

$$p(\theta) \propto \exp\left(-\frac{1}{2}(\theta-\mu_0)^T Q_0(\theta-\mu_0)\right).$$

The posterior is proportional to their product,

$$p(\theta \mid y) \propto \exp\left(-\frac{1}{2}(\theta-y)^T R(\theta-y) - \frac{1}{2}(\theta-\mu_0)^T Q_0(\theta-\mu_0)\right) = \exp\left(-\frac{1}{2}\,q(\theta)\right)$$

Here

$$\begin{aligned} q(\theta) &= (\theta-y)^T R(\theta-y) + (\theta-\mu_0)^T Q_0(\theta-\mu_0) \\ &= \theta^T R\theta - 2\theta^T Ry + y^T Ry + \theta^T Q_0\theta - 2\theta^T Q_0\mu_0 + \mu_0^T R\mu_0 \\ &= \theta^T(R+Q_0)\theta - 2\theta^T(Ry + Q_0\mu_0) + c, \end{aligned}$$

where the scalar c does not depend on θ. Comparing this result with (5.8), we see that the posterior is the multivariate normal $N_d(\mu_1, Q_1^{-1})$, where

$$Q_1 = Q_0 + R, \qquad Q_1\,\mu_1 = Q_0\,\mu_0 + Ry. \tag{5.9}$$

Again, posterior precision equals the prior precision plus the datum precision. In this manner one can identify the parameters of a multivariate normal distribution, by completing the square.

As in the univariate case, this result can be extended to several (conditionally) independent observations, and also to the case where both the mean vector and the precision matrix are (partially) unknown, when we employ an appropriate conjugate prior.

5.5.2 Linear regression when the error variance is known

The sampling model in linear regression can be described by stating that

$$Y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \dots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently of each other and independently of β, where we assume that the error variance σ^2 is known. Here x_i is the known covariate vector corresponding to the i'th response Y_i, and $\beta \in \mathbb{R}^p$ is the vector of regression coefficients.

It is more convenient to rewrite the linear regression model in matrix form, as follows,

$$[Y \mid \beta] \sim N_n(X\beta, \frac{1}{\tau} I) \tag{5.10}$$

Here X is the known $n \times p$ model matrix, which has x_i^T as its i'th row vector, and $\tau = 1/\sigma^2$ is the precision parameter of the error distribution.

If the prior is multivariate normal,

$$p(\beta) = N_p(\beta \mid \mu_0, Q_0^{-1})$$

then the posterior is also multivariate normal,

$$p(\beta \mid y) = N_p(\beta \mid \mu_1, Q_1^{-1}),$$

with

$$Q_1 = Q_0 + \tau X^T X, \qquad Q_1 \, \mu_1 = Q_0 \, \mu_0 + \tau X^T y. \tag{5.11}$$

This follows similarly as in the previous section, namely from

$$\begin{aligned} p(\beta \mid y) &\propto p(y \mid \beta) \, p(\beta) \\ &= N_n(y \mid X\beta, \frac{1}{\tau} I) \, N_p(\beta \mid \mu_0, Q_0^{-1}) \\ &\propto \exp\left(-\frac{1}{2} \tau \, (y - X\beta)^T (y - X\beta) - \frac{1}{2} (\beta - \mu_0)^T Q_0 (\beta - \mu_0) \right) \\ &= \exp\left(-\frac{1}{2} q(\beta) \right), \end{aligned}$$

where the quadratic form $q(\beta)$ is

$$\begin{aligned} q(\beta) &= \tau \, (y - X\beta)^T (y - X\beta) + (\beta - \mu_0)^T Q_0 (\beta - \mu_0) \\ &= \beta^T \left(\tau X^T X + Q_0 \right) \beta - 2 \, \beta^T (\tau X^T y + Q_0 \, \mu_0) + c_1. \end{aligned}$$

Comparing this result with (5.8), we get the previously announced formulas (5.11).

5.5.3 Known mean but unknown precision matrix

At this point we need to introduce the Wishart distribution which is a joint distribution for the distinct entries of a symmetric random $d \times d$ matrix X. When we use the parametrization of Bernanrdo and Smith [1], then the Wishart distribution with parameters $\alpha > (d-1)/2$ and B (a symmetric positive definite matrix) has the density

$$\mathrm{Wish}_d(X \mid \alpha, B) = \frac{\det(B)^\alpha}{\Gamma_d(\alpha)} \det(X)^{\alpha-(d+1)/2} \exp(-\operatorname{tr}(B\,X)), \qquad X > 0$$

Here $\Gamma_d(\cdot)$ is the generalized gamma function, see (A.6) and $\operatorname{tr}(M)$ denotes the trace of the square matrix M, i.e.,

$$\operatorname{tr}(M) = \sum_i m_{ii}.$$

Further, the qualification $X > 0$ means that the above expression applies when X is not only symmetric but also positive definite, otherwise the Wishart pdf is zero. The Wishart density is the joint pdf of the $d(d+1)/2$ distinct entries of the symmetric matrix X, e.g., the elements $x_{ij}, i \geq j$ on or below the main diagonal. When $d = 1$, then $\mathrm{Wish}_d(x \mid \alpha, \beta)$ reduces to $\mathrm{Gam}(x \mid \alpha, \beta)$.

E.g., when $d = 2$ then the symmetric matrix X can be written using only the elements x_{11}, x_{21} and x_{22} as follows,

$$X = \begin{bmatrix} x_{11} & x_{21} \\ x_{21} & x_{22} \end{bmatrix}$$

A symmetric matrix is positive definite if and only if all of its eigenvalues are positive and if and only if all of its leading principal minors are positive. In this 2×2 case, $X = [x_{ij}]$ is positive definite if and only if

$$x_{11} > 0, \qquad x_{11}x_{22} - x_{21}^2 > 0$$

Writing out the determinant and the trace in terms of the matrix elements, we obtain the following expression for the joint density of the Wishart distribution,

$$\mathrm{Wish}_2(x_{11}, x_{21}, x_{22} \mid \alpha, B) = \frac{\det(B)^\alpha}{\sqrt{\pi}\,\Gamma(\alpha)\Gamma(\alpha - \frac{1}{2})} \\ (x_{11}x_{22} - x_{21}^2)^{\alpha-\frac{3}{2}} \exp\left(-(\beta_{11}x_{11} + 2\beta_{21}x_{21} + \beta_{22}x_{22})\right),$$

if $x_{11} > 0$ and $x_{11}x_{22} - x_{21}^2 > 0$, and the pdf is zero otherwise.

Now we assume that

$$Y_i \mid Q \overset{\text{i.i.d.}}{\sim} N_d(\mu, Q^{-1}), \qquad i = 1, \ldots, n.$$

Here $\mu \in \mathbb{R}^d$ is a known vector, and $Q \in \mathbb{R}^{d \times d}$ is an unknown symmetric positive definite matrix. The prior density is $\mathrm{Wish}_d(Q \mid \alpha, B)$.

Since

$$p(y_i \mid Q) = (2\pi)^{-d/2} \det(Q)^{1/2} \exp\left(-\frac{1}{2}(y_i - \mu)^T Q (y_i - \mu)\right),$$

the likelihood is

$$p(y \mid Q) \propto \det(Q)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^T Q (y_i - \mu)\right).$$

To proceed, we need to use certain properties of the trace of a matrix.

The matrix trace is obviously a linear function, the trace of a scalar γ is $\operatorname{tr}(\gamma) = \gamma$, and what is more, trace also satisifies the identity

$$\operatorname{tr}(C\,D) = \operatorname{tr}(D\,C)$$

whenever $C\,D$ is a square matrix (the factors need not be square matrices). Therefore we can write as follows

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\mu)^T Q(y_i-\mu) &= \sum_{i=1}^{n}\operatorname{tr}\left((y_i-\mu)^T Q(y_i-\mu)\right)\\ &= \sum_{i=1}^{n}\operatorname{tr}\left((y_i-\mu)(y_i-\mu)^T Q\right)\\ &= \operatorname{tr}\left(\left[\sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T\right] Q\right).\end{aligned}$$

Let us denote

$$S = \sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T.$$

When we use the above trace identity in the likelihood and combine likelihood with the prior, we get

$$\begin{aligned}p(Q \mid y) \propto{}& \det(Q)^{\alpha-\frac{1}{2}(d+1)}\, \exp(-\operatorname{tr}(B\,X))\\ & \det(Q)^{n/2}\, \exp(-\operatorname{tr}(\frac{1}{2}S\,X))\\ ={}& \det(Q)^{\alpha+n/2-\frac{1}{2}(d+1)}\, \exp\left[-\operatorname{tr}\left((B+\frac{1}{2}S)Q\right)\right],\end{aligned}$$

which shows that the posterior is

$$\mathrm{Wish}_d(\alpha+\frac{n}{2},\; B+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T).$$

If need be, this can be translated to the traditional parametrization of the Wishart distribution with the aid of Appendix A.7.

5.6 Conditional conjugacy

In multiparameter problems it may be difficult or impossible to use conjugate priors. However, some benefits of conjugate families can be retained, if one has conditional conjugacy in the Bayesian statistical model.

Suppose we have parameter vector θ, which we partition as $\theta = (\phi, \psi)$, where the components ϕ and ψ are not necessarily scalars. The the **full conditional** (density) of ϕ in the prior distribution is defined as

$$p(\phi \mid \psi),$$

and the full conditional (density) of of ϕ in the posterior is defined as

$$p(\phi \mid \psi, y).$$

Then ϕ exhibits conditional conjugacy, if the full conditional of ϕ in the prior and and in the posterior belong to the same family of distributions.

In practice, one notices the conditional conjugacy of ϕ as follows. The prior full conditional of ϕ is

$$p(\phi \mid \psi) \propto p(\phi, \psi),$$

when we regard the joint prior as a function of ϕ. Similarly, the posterior full conditional of ϕ is

$$p(\phi \mid \psi, y) \propto p(\phi, \psi, y) = p(\phi, \psi)\, p(y \mid \phi, \psi),$$

when we regard the joint distribution $p(\phi, \psi, y)$ as a function of ϕ. If we recognize the functional forms of the prior full conditional and the posterior full conditional, then we have conditional conjugacy.

If we partition the parameter vector into k components, $\theta = (\theta_1, \ldots, \theta_k)$ (which are not necessarily scalars), then sometimes all the components are conditionally conjugate. In other cases, only some of the components turn out to be conditionally conjugate.

Example 5.1. [Conditional conjugacy in linear regression] Consider the linear regression model

$$[Y \mid \beta, \tau] \sim N_n(X\beta, \frac{1}{\tau} I), \tag{5.12}$$

where now both β and the precision of the error distribution τ are unknown. If the prior distribution is of the form

$$p(\beta, \tau) = p(\tau)\, N_p(\beta \mid \mu_0(\tau), (Q_0(\tau))^{-1}),$$

where the mean and the precision matrix of the conditional prior of β are arbitrary functions of τ, then the full conditional of β can be obtained from equations (5.11), since τ is considered known in $p(\beta \mid \tau, y)$.

On the other hand, if the prior distribution is of the form

$$p(\beta, \tau) = p(\beta)\ \mathrm{Gam}(\tau \mid a_0(\beta), b_0(\beta)),$$

then an easy calculation shows that the full conditional of τ is also a gamma distribution. △

5.7 Reparametrization

Suppose that we have formulated a Bayesian statistical model in terms of a parameter vector θ with a continuous distribution, but then want to reformulate it in terms of a new parameter vector ϕ, where there is a diffeomorphic correspondence between θ and ϕ. I.e., the correspondence

$$\phi = g(\theta) \quad \Leftrightarrow \quad \theta = h(\phi)$$

is one-to-one and continuously differentiable in both directions. What happens to the prior, likelihood and the posterior under such a reparametrization?

We get the prior of ϕ using the change of variables formula for densities:

$$f_\Phi(\phi) = f_\Theta(\theta) \left| \frac{\partial \theta}{\partial \phi} \right| = f_\Theta(h(\phi))\, |J_h(\phi)|.$$

If we know ϕ then we also know $\theta = h(\phi)$. Therefore the likelihood stays the same in that

$$f_{Y|\Phi}(y \mid \phi) = f_{Y|\Theta}(y \mid h(\phi)).$$

Finally, the posterior density changes in the same way as the prior density (by the change of variables formula), i.e.,

$$f_{\Phi|Y}(\phi \mid y) = f_{\Theta|Y}(\theta \mid y) \left| \frac{\partial \theta}{\partial \phi} \right| = f_{\Theta|Y}(h(\phi) \mid y)\, |J_h(\phi)|.$$

5.8 Improper priors

Sometimes one specifies a prior by stating that

$$p(\theta) \propto h(\theta),$$

where $h(\theta)$ is a non-negative function, whose integral is infinite

$$\int h(\theta)\, \mathrm{d}\theta = \infty.$$

Then there does not exist a constant of proportionality that will allow $p(\theta)$ to be a proper density, i.e., to integrate to one. In that case we have an **improper prior**. Notice that this is different from expressing the prior by the means of an unnormalized density h, which can be normalized to be a proper density. Sometimes we get a proper posterior, if we multiply an improper prior with the likelihood and then normalize.

For example, consider one normally distributed observation $Y \sim N(\theta, \tau^2)$ with a known variance τ^2, and take

$$p(\theta) \propto 1, \qquad \theta \in \mathbb{R}\,.$$

This prior is intended to represent complete prior ignorance about the unknown mean: all possible values are deemed equally likely. Calculating formally,

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)\, p(\theta) \propto \exp\left(-\frac{1}{2\tau^2}(y-\theta)^2 \right) \\ &\propto N(\theta \mid y, \tau^2) \end{aligned}$$

We obtain the same result in the limit, if we take $N(\mu_0, \sigma_0^2)$ as the prior and then let the prior variance σ_0^2 go to infinity.

One often uses improper priors in a location-scale model, with a location parameter μ and a scale parameter σ. Then it is conventional to take the prior of the location parameter to be uniform and to let the logarithm of the scale parameter σ have a uniform distribution and to take them to be independent in their improper prior. This translates to an improper prior of the form

$$p(\mu, \sigma) \propto \frac{1}{\sigma}, \qquad \mu \in \mathbb{R}, \sigma > 0 \tag{5.13}$$

by using (formally) the change of variables formula,

$$p(\sigma) = p(\tau) \left| \frac{\mathrm{d}\tau}{\mathrm{d}\sigma} \right| \propto \frac{1}{\sigma},$$

when $\tau = \log \sigma$ and $p(\tau) \propto 1$.

Some people use the so called Jeffreys' prior, which is designed to have a form which is invariant with respect to one-to-one reparametrizations. Also this leads typically to an improper prior. There are also other processes which attempt produce non-informative priors, which often turn out to be improper. (A prior is called non-informative, vague, diffuse or flat, if it plays a minimal role in the posterior distribution.)

Whereas the posterior derived from a proper prior is automatically proper, the posterior derived from an improper prior can be either proper or improper. Notice, however, that **an improper posterior does not make any sense**. If you do use an improper prior, it is *your* duty to check that the posterior is proper.

5.9 Summarizing the posterior

The posterior distribution gives a complete description of the uncertainty concerning the parameter after the data has been observed. If we use conjugate analysis inside a well-understood conjugate family, then we need only report the hyperparameters of the posterior. E.g., if the posterior is a multivariate normal (and the dimensionality is low) then the best summary is to give the mean and the covariance matrix of the posterior. However, in more complicated situations the functional form of the posterior may be opaque, and then we need to summarize the posterior.

If we have a univariate parameter, then the best description of the posterior is the plot of its density function. Additionally, we might want to calculate such summaries as the posterior mean the posterior variance, the posterior mode, the posterior median, and other selected posterior quantiles. If we cannot plot the density, but are able to simulate from the posterior, we can plot the histogram and calculate summaries (mean, variance, quantiles) from the simulated sample.

If we have a two-dimensional parameter, then we can still make contour plots or perspective plots of the density, but in higher dimensions such plots are not possible. One practical approach in a multiparameter situation is to summarize the one-dimensional marginal posteriors of the scalar components of the parameter.

Suppose that (after a rearrangement of the components) $\theta = (\phi, \psi)$, where ϕ is the scalar component of interest. Then the marginal posterior of ϕ is

$$p(\phi \mid y) = \int p(\phi, \psi \mid y) \, \mathrm{d}\psi$$

The indicated integration may be very difficult to perform analytically. However, if one has available a sample

$$(\phi_1, \psi_1), (\phi_2, \psi_2), ..., (\phi_N, \psi_N)$$

from the posterior of $\theta = (\phi, \psi)$, then $\phi_1, \phi_2, \ldots, \phi_N$ is a sample from the marginal posterior of ϕ. Hence we can summarize the marginal posterior of ϕ based on the sample $\phi_1, \ldots, \phi_N$.

5.10 Posterior intervals

We consider a univariate parameter θ which has a continuous distribution. One conventional summary of the posterior is a $100(1-\alpha)\%$ **posterior interval** of the parameter θ, which is any interval C in the parameter space such that

$$P(\Theta \in C \mid Y = y) = \int_C p(\theta \mid y) \, \mathrm{d}\theta = 1 - \alpha. \tag{5.14}$$

Here $0 < \alpha < 1$ is some fixed number, such that the required coverage probability is $1 - \alpha$; usual choices are $\alpha = 0.05$ or $\alpha = 0.1$ corresponding to 95 % and 90 % probability intervals, respectively. Some authors call such intervals **probability intervals**, **credible intervals** (or credibility intervals) or **Bayesian confidence intervals**.

The posterior intervals have the direct probabilistic interpretation (5.14). In contrast, the confidence intervals of frequentist statistics have probability interpretations only with reference to (hypothetical) sampling of the data under identical conditions.

Within the frequentist framework, the parameter is an unknown deterministic quantity. A frequentist confidence interval either covers or does not cover the true parameter value. A frequentist statistician constructs a frequentist confidence interval at significance level $\alpha 100\%$ in such a way that if it were possible to sample repeatedly the data under identical conditions (i.e., using the same value for the parameter), then the relative frequency of coverage in a long run of repetitions would be about $1 - \alpha$. But for the data at hand, the calculated frequentist confidence interval still either covers or does not cover the true parameter value, and we do not have guarantees for anything more. Many naive users of statistics (and even some textbook authors) mistakenly believe that their frequentist confidence intervals have the simple probability interpretation belonging to posterior intervals.

The coverage requirement (5.14) needs to be supplemented by other criteria. To demonstrate why this is so, let q be the quantile function of posterior (which may be explicitely available in simple conjugate situations and may be approximated with the empirical quantile function in other situations). Then the interval

$$[q(t), q(1 - (\alpha - t))]$$

has the required coverage probability $1 - \alpha$ for any value $0 < t < \alpha$; cf. (2.5) and (2.6).

In practice it is easiest to use the **equal tail (area) interval** (or central interval), whose end points are selected so that $\alpha/2$ of the posterior probability lies to the left and $\alpha/2$ to the right of the intervals. By the definition of the quantile function, the equal tail posterior interval is given by

$$[q(\alpha/2), q(1 - \alpha/2)]. \tag{5.15}$$

If the quantile function is not available, but one has available a sample $\theta_1, \ldots, \theta_N$ from the posterior, then one can use the empirical quantiles calculated from the sample.

Many authors recommend the **highest posterior density (HPD)** region, which is defined as the set

$$C_t = \{\theta : f_{\Theta|Y}(\theta \mid y) \geq t\},$$

where the threshold t has to be selected so that

$$P(\Theta \in C_t) = 1 - \alpha.$$

Often (but not always) the HPD region turns out to be an interval. Then it can proven to be the shortest interval with the desired coverage $100(1-\alpha)\%$. However, calculating a HPD interval is more difficult than calculating an equal tail interval.

In a multiparameter situation one usually examines one parameter at a time. Let ϕ be the scalar parameter of interest in $\theta = (\phi, \psi)$, and suppose that we have available a sample

$$(\phi_1, \psi_1), (\phi_2, \psi_2), ..., (\phi_N, \psi_N)$$

from the posterior. Then $\phi_1, \phi_2, \ldots, \phi_N$ is a sample from the marginal posterior of ϕ. Hence the central marginal posterior interval of ϕ can be calculated as in (5.15), when q is the empirical quantile function based on $\phi_1, \ldots, \phi_N$.

5.11 Literature

See, e.g., Bernardo and Smith [1] for further results on conjugate analysis. The books by Gelman *et al.* [4], Carlin and Louis [2] and O'Hagan and Forster [5] are rich sources of ideas on Bayesian modeling and analysis. Sufficiency is a central concept in parametric statistics. See, e.g., Schervish [6] for a discussion.

Bibliography

[1] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* John Wiley & Sons, 2000. First published in 1994.

[2] Bradley P. Carlin and Thomas A. Louis. *Bayesian Methods for Data Analysis.* Chapman & Hall/CRC, 3rd edition, 2009.

[3] Ronald Christensen, Wesley Johnson, Adam Branscum, and Timothy E. Hanson. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians.* Texts in Statistical Science. CRC Press, 2011.

[4] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman & Hall/CRC Press, 2nd edition, 2004.

[5] Anthony O'Hagan and Jonathan Forster. *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics.* Arnold, second edition, 2004.

[6] Mark J. Schervish. *Theory of Statistics.* Springer series in statistics. Springer-Verlag, 1995.

Chapter 6

Approximations

6.1 The grid method

When one is confronted with a low-dimensional problem with a continuous parameter, then it is usually easy to approximate the posterior density on a dense grid of points which covers the relevant part of the parameter space. We discuss the method for a one-dimensional parameter θ.

We suppose that the posterior is available in the unnormalized form

$$f_{\Theta|Y}(\theta \mid y) = \frac{1}{c(y)}\, q(\theta \mid y),$$

where we know how to evaluate the unnormalized density $q(\theta \mid y)$, but do not necessarily know the value of the normalizing constant $c(y)$.

Instead of the original parameter space, we consider a finite interval $[a, b]$, which should cover most of the mass of the posterior distribution. We divide $[a, b]$ evenly into N subintervals

$$B_i = [a + (i-1)h, a + ih], \qquad i = 1, \ldots, N.$$

The width h of one subinterval is

$$h = \frac{b-a}{N}.$$

Let θ_i be the midpoint of the i'th subinterval,

$$\theta_i = a + (i - \frac{1}{2})h, \qquad i = 1, \ldots, N.$$

We use the midpoint rule for numerical integration. This means that we approximate the integral over the i'th subinterval of any function g by the rule

$$\int_{B_i} g(\theta)\, \mathrm{d}\theta \approx h g(\theta_i). \tag{6.1}$$

Using the midpoint rule on each of the subintervals, we get the following approximation for the normalizing constant

$$\begin{aligned} c(y) &= \int q(\theta \mid y) \, \mathrm{d}\theta \approx \int_a^b q(\theta \mid y) \, \mathrm{d}\theta = \sum_{i=1}^N \int_{B_i} q(\theta \mid y) \, \mathrm{d}\theta \\ &\approx h \sum_{i=1}^N q(\theta_i \mid y) \end{aligned} \tag{6.2}$$

Using this approximation, we can approximate the value of the posterior density at the point θ_i,

$$f_{\Theta|Y}(\theta_i \mid y) = \frac{1}{c(y)} \, q(\theta_i \mid y) \approx \frac{1}{h} \frac{q(\theta_i \mid y)}{\sum_{j=1}^N q(\theta_j \mid y)}. \tag{6.3}$$

We also obtain approximations for the posterior probabilities of the subintervals,

$$\begin{aligned} P(\Theta \in B_i \mid Y = y) &= \int_{B_i} f_{\Theta|Y}(\theta \mid y) \, \mathrm{d}\theta \approx h f_{\Theta|Y}(\theta_i \mid y) \\ &\approx \frac{q(\theta_i \mid y)}{\sum_{j=1}^N q(\theta_j \mid y)}. \end{aligned} \tag{6.4}$$

By following the same reasoning which lead to (6.2), we may form the approximation

$$\int k(\theta) \, q(\theta \mid y) \, \mathrm{d}\theta \approx h \sum_{i=1} k(\theta_i) \, q(\theta_i \mid y)$$

basically for any function k such that $k(\theta) \, q(\theta \mid y)$ differs appreciably from zero only on the interval (a, b). This can be used to approximate the posterior expection of an arbitrary function $k(\theta)$ of the parameter, by

$$\begin{aligned} E(k(\Theta) \mid Y = y) &= \int k(\theta) \, f_{\Theta|Y}(\theta \mid y) \, \mathrm{d}\theta = \frac{\int k(\theta) \, q(\theta \mid y) \, \mathrm{d}\theta}{\int q(\theta \mid y) \, \mathrm{d}\theta} \\ &\approx \frac{\sum_{i=1}^N k(\theta_i) \, q(\theta_i \mid y)}{\sum_{j=1}^N q(\theta_j \mid y)} \end{aligned} \tag{6.5}$$

These approximations can be surprisingly accurate even for moderate values of N provided we are able to identify an interval $[a, b]$, which covers the essential part of posterior distribution.

To summarize, the grid method for approximating the posterior density or for or simulating from it is the following.

- First evaluate the unnormalized posterior density $q(\theta \mid y)$ at a regular grid of points $\theta_1, \ldots, \theta_N$ with spacing h. The grid should cover the main support of the posterior density.

- If you want to plot the posterior density, normalize these values by dividing by their sum and additionally by the bin width h as in eq. (6.3). This gives an approximation to the posterior ordinates $p(\theta_i \mid y)$ at the grid points θ_i.

- If you want a sample from the posterior, sample with replacement from the grid points θ_i with probabilities proportional to the numbers $q(\theta_i \mid y)$, cf. (6.4).

- If you want to approximate the posterior expectation $E[k(\theta) \mid y]$, calculate the weighted average of the values $k(\theta_i)$ using the values $q(\theta_i \mid y)$ as weights, cf. eq. (6.5).

The midpoint rule is considered a rather crude method of numerical integration. In the numerical analysis literature, there are available much more sophisticated methods of numerical integration (or numerical quadrature) and they can be used in a similar manner. Besides dimension one, these kinds of approaches can be used in dimensions two or three. However, as the dimensionality of the parameter space grows, computing at every point in a dense multidimensional grid becomes more and more expensive.

6.2 Normal approximation to the posterior

We now try to approximate a posterior density by a normal density based on the behavior of the posterior density at its mode. This approximation can be quite accurate, when the sample sizes is large, provided the posterior is unimodal. We will call the resulting approximation a normal approximation to the posterior, but the result is sometimes also called a Laplace approximation or a modal approximation. A normal approximation can be used directly as an approximate description of the posterior. However, such an approximation can be utilized also indirectly, e.g., to form a good proposal distribution for the Metropolis–Hastings method.

We first discuss normal approximation in the univariate situation. The statistical model has a single parameter θ, which has a continuous distribution. We do know an unnormalized version $q(\theta \mid y)$ of the posterior density, but the normalizing constant is usually unknown. We consider the case, where $\theta \mapsto q(\theta \mid y)$ is unimodal: i.e., it has only one local maximum. We suppose that we have located the mode $\hat{\theta}$ of the unnormalized posterior $q(\theta \mid y)$. Notice that $\hat{\theta}$ is also the posterior mode, which is also called the MAP (maximum a posteriori) estimate. Actually, $\hat{\theta}$ depends on the data y, but we suppress this dependence in our notation. Usually we would have to run some numerical optimization algorithm in order to find the mode.

The basic idea of the method is to use the second degree Taylor polynomial of the log-posterior (the logarithm of the posterior density) centered on the mode $\hat{\theta}$,

$$\log f_{\Theta|Y}(\theta \mid y) \approx \log f_{\Theta|Y}(\hat{\theta} \mid y) + b(\theta - \hat{\theta}) - \frac{1}{2} A(\theta - \hat{\theta})^2, \tag{6.6}$$

where

$$b = \frac{\partial}{\partial \theta} \log f_{\Theta|Y}(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = \frac{\partial}{\partial \theta} \log q(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = 0,$$

and

$$A = -\frac{\partial^2}{\partial \theta^2} \log f_{\Theta|Y}(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = -\frac{\partial^2}{\partial \theta^2} \log q(\theta \mid y)\bigg|_{\theta=\hat{\theta}}.$$

Notice the following points.

- The first and higher order (partial) derivatives with respect to θ of $\log q(\theta \mid y)$ and $\log f_{\Theta|Y}(\theta \mid y)$ agree, since these function differ only by an additive constant (which depends on y but not on θ).

- The first order term of the Taylor expansion disappears, since $\hat{\theta}$ is also the mode of the log-posterior $\log f_{\Theta|Y}(\theta \mid y)$.
- $A \geq 0$, since $\hat{\theta}$ is a maximum of $q(\theta \mid y)$. For the following, we need to assume that $A > 0$.

Taking the exponential of the second degree Taylor approximation (6.6), we see that we may approximate the posterior by the function

$$\pi_{\text{approx}}(\theta) \propto \exp\left(-\frac{A}{2}(\theta - \hat{\theta})^2\right),$$

at least in the vicinity of the mode $\hat{\theta}$. Luckily, we recognize that $\pi_{\text{approx}}(\theta)$ is an unnormalized form of the density of the normal distribution with mean $\hat{\theta}$ and variance $1/A$. The end result is that the posterior distribution can be approximated with the normal distribution

$$N\left(\hat{\theta}, \frac{1}{-L''(\hat{\theta})}\right), \tag{6.7}$$

where $L(\theta)$ is the logarithm of the unnormalized posterior,

$$L(\theta) = \log q(\theta \mid y)$$

and $L''(\hat{\theta})$ is the second derivative of $L(\theta)$ evaluated at the mode $\hat{\theta}$.

The multivariate analog of the result starts with the second degree expansion of the log-posterior centered on its mode $\hat{\theta}$,

$$\log f_{\Theta|Y}(\theta \mid y) \approx \log f_{\Theta|Y}(\hat{\theta} \mid y) + 0 - \frac{1}{2}(\theta - \hat{\theta})^T A(\theta - \hat{\theta}),$$

where A is the negative Hessian matrix of $L(\theta) = \log q(\theta \mid y)$ evaluated at the mode,

$$A_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f_{\Theta|Y}(\theta \mid y)\bigg|_{\theta=\hat{\theta}} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} L(\theta)\bigg|_{\theta=\hat{\theta}} = -\left[\frac{\partial^2}{\partial\theta\, \partial\theta^T} L(\theta)\bigg|_{\theta=\hat{\theta}}\right]_{ij}$$

The first degree term of the expansion vanishes, since $\hat{\theta}$ is the mode of the log-posterior. Here A is at least positively semidefinite, since $\hat{\theta}$ is a maximum. If A is positively definite, we can proceed with the normal approximation.

Exponentiating, we find out that approximately (at least near the mode)

$$f_{\Theta|Y}(\theta \mid y) \propto \exp\left(-\frac{1}{2}(\theta - \hat{\theta})^T A(\theta - \hat{\theta})\right).$$

Therefore we can approximate the posterior with the corresponding multivariate normal distribution with mean $\hat{\theta}$ and covariance matrix given by A^{-1}, i.e., the approximating normal distribution is

$$N\left(\hat{\theta}, \left(-L''(\hat{\theta})\right)^{-1}\right), \tag{6.8}$$

where $L''(\hat{\theta})$ is the Hessian matrix of the logarithm of the unnormalized posterior, $L(\theta) = \log q(\theta \mid y)$, evaluated at its mode $\hat{\theta}$. The precision matrix of the approximating normal distribution is the negative Hessian of the log-posterior evaluated at the posterior mode. Another characterization for the precision matrix is that it is the Hessian of the negative log-posterior evaluated at the posterior mode. The covariance matrix of the normal approximation is the inverse of its precision matrix.

Typically the mode of the log-posterior (or the maximum point of the negative log-posterior) would be calculated using some numerical optimization algorithm. The Hessian would then be calculated using numerical differentiation, see Sec. B.7 for an example.

Before using the normal approximation, it is often advisable to reparameterize the model so that the transformed parameters are defined on the whole real line and have roughly symmetric distributions. E.g., one can use logarithms of positive parameters and apply the logit function to parameters which take values on the interval $(0, 1)$. The normal approximation is then constructed for the transformed parameters, and the approximation can then be translated back to the original parameter space. One must, however, remember to multiply by the appropriate Jacobians.

Example 6.1. We consider the unnormalized posterior

$$q(\theta \mid y) = \theta^{y_4} \, (1-\theta)^{y_2+y_3} \, (2+\theta)^{y_1}, \qquad 0 < \theta < 1,$$

where $y = (y_1, y_2, y_3, y_4) = (13, 1, 2, 3)$. The mode and the second derivative of $L(\theta) = \log q(\theta \mid y)$ evaluated at the mode are given by

$$\hat{\theta} \approx 0.677, \qquad L''(\hat{\theta}) \approx -37.113.$$

(The mode $\hat{\theta}$ can be found by solving a quadratic equation.) The resulting normal approximation in the original parameter space is $N(0.677, 1/37.113)$.

We next reparametrize by defining ϕ as the logit of θ,

$$\phi = \operatorname{logit}(\theta) = \ln \frac{\theta}{1-\theta} \quad \Leftrightarrow \quad \theta = \frac{\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}}.$$

The given unnormalized posterior for θ transforms to the following unnormalized posterior for ϕ,

$$\begin{aligned} \tilde{q}(\phi \mid y) &= q(\theta \mid y) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\phi} \right| \\ &= \left(\frac{\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}} \right)^{y_4} \left(\frac{1}{1+\mathrm{e}^{\phi}} \right)^{y_2+y_3} \left(\frac{2+3\mathrm{e}^{\phi}}{1+\mathrm{e}^{\phi}} \right)^{y_1} \frac{\mathrm{e}^{\phi}}{(1+\mathrm{e}^{\phi})^2}. \end{aligned}$$

The mode and the second derivative of $\tilde{L}(\phi) = \log \tilde{q}(\phi \mid y)$ evaluated at the mode are given by

$$\hat{\phi} \approx 0.582, \qquad \tilde{L}''(\hat{\phi}) \approx -2.259.$$

(Also $\hat{\phi}$ can be found by solving a quadratic.) This results in the normal approximation $N(0.582, 1/2.259)$ for the logit of θ.

When we translate that approximation back to the original parameter space, we get the approximation

$$f_{\Theta|Y}(\theta \mid y) \approx N(\phi \mid 0.582, 1/2.259) \left| \frac{\mathrm{d}\phi}{\mathrm{d}\theta} \right|,$$

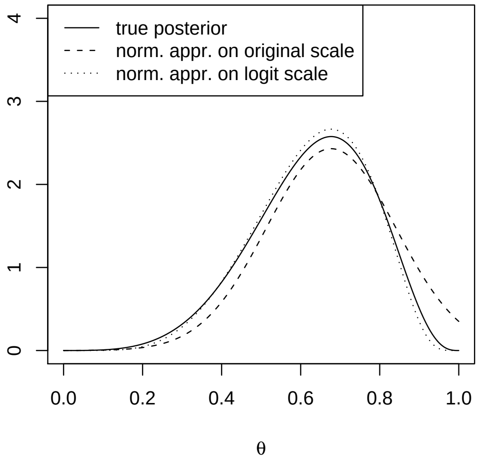


Figure 6.1: The exact posterior density (solid line) together with its normal approximation (dashed line) and the approximation based on the normal approximation for the logit of θ. The last approximation is markedly non-normal on the original scale, and it is able to capture the skewness of the true posterior density.

i.e.,

$$f_{\Theta|Y}(\theta \mid y) \approx N(\operatorname{logit}(\theta) \mid 0.582, 1/2.259)\, \frac{1}{\theta(1-\theta)}.$$

Both of these approximations are plotted in Figure 6.1 together with the true posterior density (whose normalizing constant can be found exactly). △

6.3 Connection to the traditional frequentist asymptotics

Here we discus the relationship of the normal approximation (6.8) to the frequentist asymptotics of the maximum likelihood estimator. The unnormalized version of the posterior density is of the form

$$q(\theta \mid y) = k(y)\, f_{Y|\Theta}(y \mid \theta)\, f_\Theta(\theta) = k(y)\, p(y \mid \theta)\, p(\theta),$$

where $p(\theta)$ is the prior, $p(y \mid \theta)$ is the likelihood, and $k(y)$ is any convenient constant which may depend on the data but not on the parameter vector. Therefore the logarithm of the unnormalized posterior is

$$L(\theta) = \log q(\theta \mid y) = \log k(y) + \ell(\theta) + \log p(\theta),$$

where $\ell(\theta) = p(y \mid \theta)$ is the log-likelihood. Therefore the negative Hessian of $L(\theta)$ is

$$-L''(\theta) = -\ell''(\theta) - \frac{\partial^2}{\partial\theta\, \partial\theta^T} \log p(\theta)$$

Here the negative Hessian of the log-likelihood is called the **observed (Fisher) information (matrix)**, and we denote it by $J(\theta)$,

$$J(\theta) = -\ell''(\theta) = -\frac{\partial^2}{\partial\theta\,\partial\theta^T} \log p(y \mid \theta). \tag{6.9}$$

The negative Hessian of the log-posterior equals the sum of the observed information and the negative Hessian of the log-prior.

If the sample size is large, then the likelihood dominates the prior in the sense that the likelihood is highly peaked while the prior is relatively flat in the region where the posterior density is appreciable. In large samples the mode of the log-posterior $\hat{\theta}$ and the mode of the log-likelihood (the maximum likelihood estimator, MLE) $\hat{\theta}_{\mathrm{MLE}}$ are approximately equal, and also the Hessian matrix of the log-posterior is approximately the same as the Hessian of the log-likelihood. Combining these two approximations, we get

$$\hat{\theta} \approx \hat{\theta}_{\mathrm{MLE}}, \qquad -L''(\hat{\theta}) \approx J(\hat{\theta}_{\mathrm{MLE}}).$$

When we plug these approximations in the normal approximation (6.8), we see that in large samples the posterior is approximately normal with mean equal to the MLE and covariance matrix given by the inverse of the observed information,

$$p(\theta \mid y) \approx N\left(\theta \mid \hat{\theta}_{\mathrm{MLE}}, [J(\hat{\theta}_{\mathrm{MLE}})]^{-1}\right). \tag{6.10}$$

This approximation should be compared with the well-known frequentist asymptotic distribution results for the maximum likelihood estimator. Loosely, these results can be summarized so that the sampling distribution of the maximum likelihood estimator is asymptotically normal with mean equal to the MLE and covariance matrix equal to the inverse of the observed information. In order to write this approximation as a formula, we need to indicate the dependence of the maximum likelihood estimator on the data as follows,

$$\hat{\theta}_{\mathrm{MLE}}(Y) \overset{\mathrm{d}}{\approx} N\left(\hat{\theta}_{\mathrm{MLE}}(y), [J(\hat{\theta}_{\mathrm{MLE}}(y))]^{-1}\right). \tag{6.11}$$

Here Y is a random vector from the sampling distribution of the data, and so $\hat{\theta}_{\mathrm{MLE}}(Y)$ is the maximum likelihood estimator considered as a random variable (or random vector). In contrast, $\hat{\theta}_{\mathrm{MLE}}(y)$ is the maximum likelihood estimate calculated from the observed data y.

Comparing equations (6.10) and (6.11) we see that for large samples the posterior distribution can be approximated using the same formulas that (frequentist) statisticians use for the maximum likelihood estimator. In large samples the influence of the prior vanishes, and then one does need to spend much effort on formulating the prior distribution so that it would reflect all available prior information. However, in small samples careful formulation of the prior is important.

6.4 Posterior expectations using Laplace approximation

Laplace showed in the 1770's how one can form approximations to integrals of highly peaked positive functions by integrating analytically a suitable normal approximation. We will now apply this idea to build approximations to posterior expectations. We assume that the posterior density is highly peaked while the function k, whose posterior expectation we seek is relatively flat. The posterior density is typically known only in the unnormalized form $q(\theta \mid y)$, and then

$$E[k(\Theta) \mid Y = y] = \frac{\int k(\theta)\, q(\theta \mid y)\, \mathrm{d}\theta}{\int q(\theta \mid y)\, \mathrm{d}\theta}. \tag{6.12}$$

Tierney and Kadane [4] approximated separately the numerator and the denominator of eq. (6.12) using Laplace's method, and analyzed the resulting error.

To introduce the idea of Laplace's approximation (or Laplace's method), consider a highly peaked function $L(\theta)$ of a scalar variable θ such that $L(\theta)$ has a unique mode (i.e., a maximum) at $\hat{\theta}$. Suppose that $g(\theta)$ is a function, which varies slowly. We seek an approximation to the integral

$$I = \int g(\theta)\, \mathrm{e}^{L(\theta)}\, \mathrm{d}\theta. \tag{6.13}$$

Heuristically, the integrand is negligible when we go far away from $\hat{\theta}$, and so we should be able to approximate the integral I by a simpler integral, where we take into account only the local behavior of $L(\theta)$ around its mode. To this end, we first approximate $L(\theta)$ by its second degree Taylor polynomial centered at the mode $\hat{\theta}$,

$$L(\theta) \approx L(\hat{\theta}) + 0 \cdot (\theta - \hat{\theta}) + \frac{1}{2} L''(\hat{\theta})(\theta - \hat{\theta})^2.$$

Since $g(\theta)$ is slowly varying, we may approximate the integrand as follows

$$g(\theta)\, \mathrm{e}^{L(\theta)} \approx g(\hat{\theta}) \, \exp\left(L(\hat{\theta}) - \frac{1}{2} Q (\theta - \hat{\theta})^2) \right),$$

where

$$Q = -L''(\hat{\theta}).$$

For the following, we must assume that $L''(\hat{\theta}) < 0$. Integrating the approximation, we obtain

$$\begin{aligned} I &\approx \int g(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})} \, \exp(-\frac{1}{2} Q (\theta - \hat{\theta})^2)\, \mathrm{d}\theta \\ &= \frac{\sqrt{2\pi}}{\sqrt{Q}}\, g(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})} \end{aligned} \tag{6.14}$$

This is the univariate case of Laplace's approximation. (Actually, it is just the leading term in a Laplace expansion, which is an asymptotic expansion for the integral.)

To handle the multivariate result, we use the normalizing constant of the $N_d(\mu, Q^{-1})$ distribution to evaluate the integral

$$\int \exp\left(-\frac{1}{2} (x - \mu)^T Q (x - \mu) \right) \mathrm{d}x = \frac{(2\pi)^{d/2}}{\sqrt{\det Q}}. \tag{6.15}$$

This result is valid for any symmetric and positive definite $d \times d$ matrix Q. Integrating the multivariate second degree approximation of $g(\theta) \exp(L(\theta))$, we obtain

$$I = \int g(\theta)\, \mathrm{e}^{L(\theta)}\, \mathrm{d}\theta \approx \frac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, g(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})}, \tag{6.16}$$

where d is the dimensionality of θ, and Q is the negative Hessian of L evaluated at the mode,

$$Q = -L''(\hat{\theta}),$$

and we must assume that the $d \times d$ matrix Q is positively definite.

Using these tools, we can approximate the posterior expectation of $k(\theta)$ (see (6.12)) in several different ways. One idea is to approximate the numerator by choosing

$$g(\theta) = k(\theta), \qquad \mathrm{e}^{L(\theta)} = q(\theta \mid y)$$

in eq. (6.16), and then to approximate the denominator by choosing

$$g(\theta) \equiv 1, \qquad \mathrm{e}^{L(\theta)} = q(\theta \mid y).$$

These choices yield the approximation

$$E[k(\Theta) \mid Y = y] \approx \frac{\dfrac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, k(\hat{\theta})\, \mathrm{e}^{L(\hat{\theta})}}{\dfrac{(2\pi)^{d/2}}{\sqrt{\det(Q)}}\, \mathrm{e}^{L(\hat{\theta})}} = k(\hat{\theta}), \tag{6.17}$$

where

$$\hat{\theta} = \arg\max L(\theta), \qquad Q = -L''(\hat{\theta}).$$

Here we need a single maximization, and do not need to evaluate the Hessian at all.

A less obvious approach is to choose

$$g(\theta) \equiv 1, \qquad \mathrm{e}^{L(\theta)} = k(\theta)\, q(\theta \mid y)$$

to approximate the numerator, and

$$g(\theta) \equiv 1, \qquad \mathrm{e}^{L(\theta)} = q(\theta \mid y)$$

to approximate the denominator. Here we need to assume that k is a positive function, i.e., $k > 0$. The resulting approximation is

$$E[k(\Theta) \mid Y = y] \approx \left(\frac{\det(Q)}{\det(Q^*)} \right)^{1/2} \frac{k(\hat{\theta}^*)\, q(\hat{\theta}^* \mid y)}{q(\hat{\theta} \mid y)}, \tag{6.18}$$

where

$$\hat{\theta}^* = \arg\max[k(\theta)\, q(\theta \mid y)], \qquad \hat{\theta} = \arg\max q(\theta \mid y).$$

and Q^* and Q are the negative Hessians

$$Q^* = -L^{*''}(\hat{\theta}^*), \qquad Q = -L''(\hat{\theta}),$$

where

$$L^*(\theta) = \log(k(\theta)\, q(\theta \mid y)), \qquad L(\theta) = \log q(\theta \mid y).$$

We need two separate maximizations and need to evaluate two Hessians for this approximation.

Tierney and Kadane analyzed the errors committed in these approximations in the situation, where we have n (conditionally) i.i.d. observations, and the sample size n grows. The first approximation (6.17) has relative error of order $O(n^{-1})$, while the second approximation (6.18) has relative error of order $O(n^{-2})$. That is,

$$E[k(\Theta) \mid Y = y] = k(\hat{\theta})\ \left(1 + O(n^{-1})\right)$$

and

$$E[k(\Theta) \mid Y = y] = \left(\frac{\det(Q)}{\det(Q^*)}\right)^{1/2} \frac{k(\hat{\theta}^*)\, q(\hat{\theta}^* \mid y)}{q(\hat{\theta} \mid y)}\ \left(1 + O(n^{-2})\right).$$

Hence the second approximation is much more accurate (at least asymptotically).

6.5 Posterior marginals using Laplace approximation

Tierney and Kadane discuss also an approximation to the marginal posterior, when the parameter vector θ is composed of two vector components $\theta = (\phi, \psi)$. The form of the approximation is easy to derive, and was earlier discussed by Leonard [1]. However, Tierney and Kadane [4, Sec. 4] were the first to analyze the error in this Laplace approximation. We first derive the form of the approximation, and then make some comments on the error terms based on the discussion of Tierney and Kadane.

Let $q(\phi, \psi \mid y)$ be an unnormalized form of the posterior density, based on which we try to approximate the normalized marginal posterior $p(\phi \mid y)$. Let the dimensions of ϕ and ψ be d_1 and d_2, respectively. We have

$$p(\phi \mid y) = \int p(\phi, \psi \mid y)\, \mathrm{d}\psi = \int \exp(\log p(\phi, \psi \mid y))\, \mathrm{d}\psi,$$

where $p(\phi, \psi \mid y)$ is the normalized posterior. The main difference with approximating a posterior expectation is the fact, that now we are integrating only over the component(s) ψ of $\theta = (\phi, \psi)$.

Fix the value of ϕ for the moment. Let $\psi^*(\phi)$ be the maximizer of the function

$$\psi \mapsto \log p(\phi, \psi \mid y),$$

and let $Q(\phi)$ be the negative Hessian matrix of this function evaluated at $\psi = \psi^*(\phi)$. Notice that we can equally well calculate $\psi^*(\phi)$ and $Q(\phi)$ as the maximizer and the negative of the $d_2 \times d_2$ Hessian matrix of $\psi \mapsto \log q(\phi, \psi \mid y)$, respectively,

$$\psi^*(\phi) = \arg\max_{\psi} (\log q(\phi, \psi \mid y)) = \arg\max_{\psi} q(\phi, \psi \mid y) \tag{6.19}$$

$$Q(\phi) = -\left[\frac{\partial^2}{\partial\psi\, \partial\psi^T} \log q(\phi, \psi \mid y)\right]_{|\psi = \psi^*(\phi)}. \tag{6.20}$$

For fixed ϕ, we have the second degree Taylor approximation in ψ,

$$\log p(\phi, \psi \mid y) \approx \log p(\phi, \psi^*(\phi) \mid y) - \frac{1}{2}(\psi - \psi^*(\phi))^T Q(\phi)(\psi - \psi^*(\phi)), \tag{6.21}$$

and we assume that matrix $Q(\phi)$ is positive definite.

Next we integrate the exponential function of the approximation (6.21) with respect to ψ, with the result

$$p(\phi \mid y) \approx p(\phi, \psi^*(\phi) \mid y) \, (2\pi)^{d_2/2} \, (\det Q(\phi))^{-1/2}.$$

To evaluate this approximation, we need the normalizing constant of the unnormalized posterior $q(\phi, \psi \mid y)$, which we obtain by another Laplace approximation, and the end result is

$$p(\phi \mid y) \approx (2\pi)^{-d_1/2} \, q(\phi, \psi^*(\phi) \mid y) \, \sqrt{\frac{\det Q}{\det Q(\phi)}}, \tag{6.22}$$

where Q is negative of the $(d_1 + d_2) \times (d_1 + d_2)$ Hessian of the function

$$(\phi, \psi) \mapsto \log q(\phi, \psi \mid y)$$

evaluated at the MAP, the maximum point of the same function. However, it is often enough to approximate the functional form of the marginal posterior. When considered as a function of ϕ, we have, approximately,

$$p(\phi \mid y) \propto q(\phi, \psi^*(\phi) \mid y) \, (\det Q(\phi))^{-1/2}. \tag{6.23}$$

The unnormalized Laplace approximation (6.23) can be given another interpretation (see, e.g., [2, 3]). By the multiplication rule,

$$p(\phi \mid y) = \frac{p(\phi, \psi \mid y)}{p(\psi \mid \phi, y)} \propto \frac{q(\phi, \psi \mid y)}{p(\psi \mid \phi, y)}.$$

This result is valid for any choice of ψ. Let us now form a normal approximation for the denominator for a fixed value of ϕ, i.e.,

$$p(\psi \mid \phi, y) \approx N(\psi \mid \psi^*(\phi), Q(\phi)^{-1}).$$

However, this approximation is accurate only in the vicinity of the mode $\psi^*(\phi)$, so let us use it only at the mode. The end result is the following approximation,

$$\begin{aligned} p(\phi \mid y) &\propto \left[\frac{q(\phi, \psi \mid y)}{N(\psi \mid \psi^*(\phi), Q(\phi)^{-1})} \right]_{|\psi = \psi^*(\phi)} \\ &= (2\pi)^{d_2/2} \det(Q(\phi))^{-1/2} \, q(\phi, \psi^*(\phi) \mid y) \\ &\propto q(\phi, \psi^*(\phi) \mid y) \, (\det Q(\phi))^{-1/2}, \end{aligned}$$

which is the same as the unnormalized Laplace approximation (6.23) to the marginal posterior of ϕ.

Tierney and Kadane show that the relative error in the approximation (6.22) is of the order $O(n^{-1})$, when we have n (conditionally) i.i.d. observations, and that most of the error comes from approximating the normalizing constant. They argue that the approximation (6.23) captures the correct functional form of the marginal posterior with relative error $O(n^{-3/2})$ and recommend that one should therefore use the unnormalized approximation (6.23), which can then be normalized by numerical integration, if need be. For instance, if we want to simulate from the approximate marginal posterior, then we can use the unnormalized approximation (6.23) directly, together with accept–reject, SIR or the grid-based simulation method of Sec. 6.1. See the articles by H. Rue and coworkers [2, 3] for imaginative applications of these ideas.

Another possibility for approximating the marginal posterior would be to build a normal approximation to the joint posterior, and then marginalize. However, a normal approximation to the marginal posterior would only give the correct result with absolute error of order $O(n^{-1/2})$, so the accuracies of both of the Laplace approximations are much better. Since the Laplace approximations yield good relative instead of absolute error, the Laplace approximations maintain good accuracy also in the tails of the densities. In contrast, the normal approximation is accurate only in the vicinity of the mode.

Example 6.2. Consider normal observations

$$[Y_i \mid \mu, \tau] \overset{\text{i.i.d.}}{\sim} N(\mu, \frac{1}{\tau}), \qquad i = 1, \ldots, n,$$

together with the non-conjugated prior

$$p(\mu, \tau) = p(\mu)\, p(\tau) = N(\mu \mid \mu_0, \frac{1}{\psi_0})\ \mathrm{Gam}(\tau \mid a_0, b_0).$$

The full conditional of μ is readily available,

$$p(\mu \mid \tau, y) = N(\mu \mid \mu_1, \frac{1}{\psi_1})$$

where

$$\psi_1 = \psi_0 + n\tau \qquad \psi_1\, \mu_1 = \psi_0\, \mu_0 + \tau \sum_{i=1}^{n} y_i$$

The mode of the full conditional $p(\mu \mid \tau, y)$ is

$$\mu^*(\tau) = \mu_1 = \frac{\psi_0\, \mu_0 + \tau \sum_{i=1}^{n} y_i}{\psi_0 + n\tau}.$$

We now use this knowledge to build a Laplace approximation to the marginal posterior of τ.

Since, as a function of μ,

$$p(\mu, \tau \mid y) \propto p(\mu \mid \tau, y),$$

$\mu^*(\tau)$ is also the mode of $p(\mu, \tau \mid y)$ for any τ. We also need the second derivative

$$\frac{\partial^2}{\partial \mu^2} (\log p(\mu, \tau \mid y)) = \frac{\partial^2}{\partial \mu^2} (\log p(\mu \mid \tau, y)) = -\psi_1,$$

for $\mu = \mu^*(\tau)$, but the derivative does not in this case depend on the value of μ at all. An unnormalized form of the Laplace approximation to the marginal posterior of τ is therefore

$$p(\tau \mid y) \propto \frac{q(\mu^*(\tau), \tau \mid y)}{\sqrt{\psi_1}}, \quad \text{where} \quad q(\mu, \tau \mid y) = p(y \mid \mu, \tau)\, p(\mu)\, p(\tau).$$

In this toy example, the Laplace approximation (6.23) for the functional form of the marginal posterior $p(\tau \mid \mu)$ is exact, since by the multiplication rule,

$$p(\tau \mid y) = \frac{p(\mu, \tau \mid y)}{p(\mu \mid \tau, y)}$$

for any choice of μ, in particular for $\mu = \mu^*(\tau)$. Here the numerator is known only in an unnormalized form.

Figure 6.2 (a) illustrates the result using data $y = (-1.4, -1.6, -2.4, 0.7, 0.6)$ and hyperparameters $\mu_0 = 0$, $\psi_0 = 0.5$, $a_0 = 1$, $b_0 = 0.1$. The unnormalized (approximate) marginal posterior has been drawn using the grid method of Sec. 6.1. Figure 6.2 (b) shows an i.i.d. sample drawn from the approximate posterior

$$\tilde{p}(\tau \mid y) \, p(\mu \mid \tau, y),$$

where $\tilde{p}(\tau \mid y)$ is a histogram approximation to the true marginal posterior $p(\tau \mid y)$, which has been sampled using the grid method.

△

Bibliography

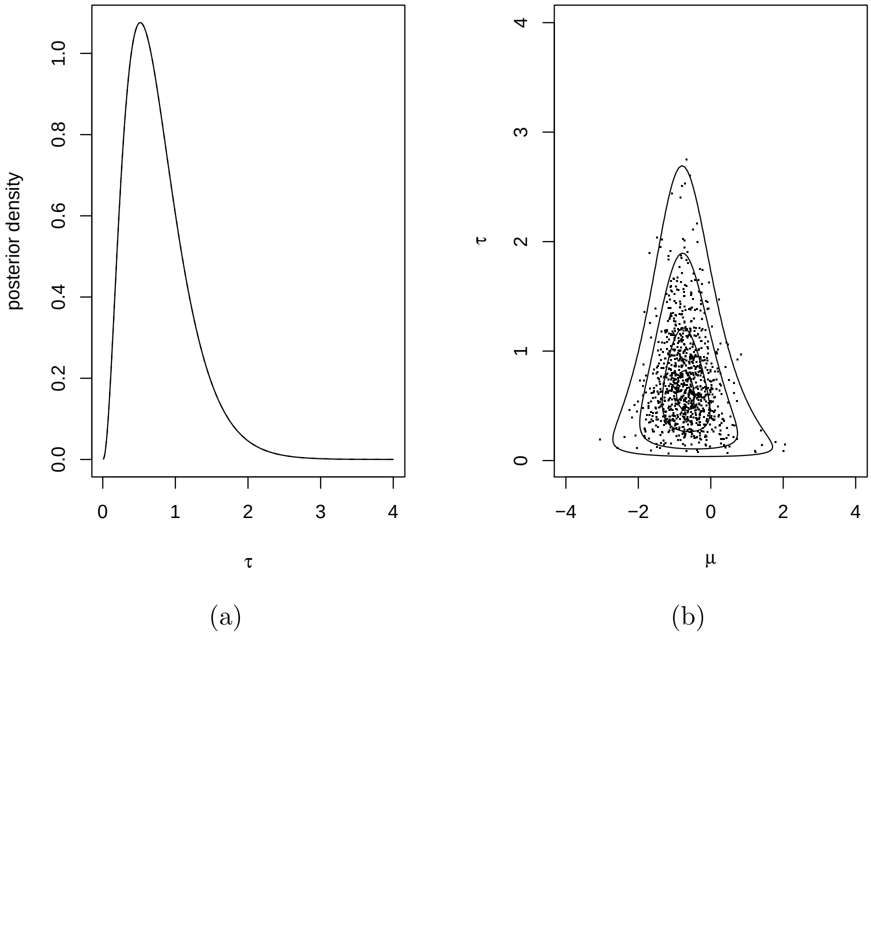
[1] Tom Leonard. A simple predictive density function: Comment. *Journal of the American Statistical Association*, 77:657–658, 1982.

[2] H. Rue and S. Martino. Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of Statistical Planning and Inference*, 137(10):3177–3192, 2007.

[3] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using integrated nested Lapalce approximations. *Journal of the Royal Statistical Society: Series B*, 2009. to appear.

[4] Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.

Figure 6.2: (a) Marginal posterior density of τ and (b) a sample drawn from the approximate joint posterior together with contours of the true joint posterior density.


posterior density
τ
(a)
τ
μ
(b)