4 Posterior Distributions

4.1 Posterior Summaries

In a Bayesian analysis, the posterior distribution encapsulates beliefs which have been updated in the light of the data. Various natural summaries of the posterior are available.

Point Estimates

Three natural summaries, as "point estimates" of the posterior are its mean, median, and mode. The mode of the posterior, which intuitively gives the "most likely" value of θ, given the data, is sometimes called the maximum Bayesian likelihood estimator.

The usual measures of the spread of a distribution (interquartile range, variance, etc.) have natural interpretations in terms of the uncertainty associated with point estimate summaries of the posterior.

Interval Estimates

Definition: If π is the density for a parameter $\theta \in \Theta$, a region $C \subseteq \Theta$ with the property that

$$\int_C \pi(\theta) d\theta = 1 - \alpha$$

is said to be a $100(1-\alpha)\%$ *credible region* for θ with respect to π. If $\Theta \subseteq \mathbf{R}$ and the set C is connected, it is called a *credible interval*.

When π is the posterior $\pi(\theta \mid x)$, credible regions and credible intervals are the Bayesian analogues of classical confidence intervals (except that they enjoy the natural intuitive interpretations).

When describing credible regions, we will assume unless it is stated otherwise, that these are with respect to the posterior $\pi(\theta \mid x)$.

Note that (as with confidence intervals) for fixed α (and continuous distribution for θ) there is not a unique $100(1-\alpha)\%$ credible region. In many contexts it is natural to report the smallest $100(1-\alpha)\%$ credible region. (This is rather less natural if the smallest such region is not connected.)

Definition: A region $C \subseteq \Theta$ is said to be a $100(1-\alpha)\%$ *highest probability density region* (HPD) with respect to π if

1. $\int_C \pi(\theta)d\theta = 1 - \alpha$
2. $\pi(\theta_1) \geq \pi(\theta_2)$ for all $\theta_1 \in C, \theta_2 \notin C$, except possibly for a subset of Θ having π probability zero.

It is straightforward to show that amongst $100(1-\alpha)\%$ credbile regions, the $100(1-\alpha)\%$ HPD has the minimum volume.

Point or interval summaries of the posterior may be useful in complicated (e.g. high dimensional) problems. The posterior distribution itself is more informative, and would be preferred, at least in simple problems.

Asymptotics

Consider a model with a prior $\pi(\theta)$ for θ and conditional on θ, data $x_1, \ldots, x_n$ i.i.d. with density $f(x \mid \theta)$. Just as in classical statistics, it is possible to examine the asymptotic behaviour of Bayesian methods when the sample size n is large.

We will describe several results informally. For a fuller treatment, see for example Bernardo and Smith, §5.3. The derivations are similar to those for asymptotic properties of maximum likelihood methods.

When the sample size n is large, it can be shown that (under suitable regularity conditions) the posterior is approximately normal. In fact, the intuition that for large n, the information in the data "swamps" the prior is correct, and the posterior will resemble a normal distribution with mean the m.l.e. $\hat{\theta}$ and the variance covariance matrix $(nI(\hat{\theta}))^{-1}$, where $I(\cdot)$ is the Fisher Information. Note that the preceding result requires that the prior density is non-zero in a region surrounding the m.l.e.

In particular, since the m.l.e. is consistent, if we generate i.i.d. data $x_1, \ldots, x_n$ from $f(x \mid \theta_0)$ for some fixed $\theta_0 \in \Theta$, then provided the prior density is nonzero around θ_0, the posterior will become more and more concentrated on θ_0 as $n \to \infty$. In this sense, all Bayesian analyses are automatically consistent.

In practice, if, for large n, the prior were zero in the region of the m.l.e., then one should reconsider the prior, or the sampling distribution, or both.

4.2 Computation

In general the problem in Bayesian Inference can be reduced to computing integrals over the posterior distribution of a parameter θ given data x.

For example

$$\Pr(\theta \in A|x) = \int I(\theta \in A)p(\theta|x)d\theta,$$

or

$$E(\theta|x) = \int \theta p(\theta|x)d\theta.$$

More generally, we are interested in approximating integrals of the form

$$E(f(\theta)|x) = \int f(\theta)p(\theta|x)d\theta.$$

4.3 Discrete case

If θ is discrete, with not too many states (e.g. $< 10^6$) then it is usually pretty straightforward to compute the posterior distribution directly, simply computing prior $\times$ likelihood and then normalizing it to get the posterior.

Integrals of interest are then, of course, simply sums.

4.4 Standard Monte Carlo

If state spaces are larger, or continuous, then we often need to approximate integrals numerically or via simulation.

If the integral is low-dimensional then direct numerical approximations (e.g. via quadrature) are often effective. In higher dimensions we usually turn to simulation. You should usually try to use the simplest approach that will work in any given situation. Here we consider methods in increasing order of complexity.

If we can simulate i.i.d. samples $\theta^{(1)}, \ldots, \theta^{(M)} \sim p(\theta|x)$ then we can approximate integrals of interest by the Monte Carlo estimate

$$E(f(\theta)|x) \approx (1/M) \sum_{m=1}^{M} f(\theta^{(m)}).$$

By the law of large numbers, this sum converges to the required expectation (assuming this is finite), as $M \to \infty$.

Standard Monte Carlo can fail for (at least) two reasons. i) It may not be straightforward to sample from the posterior distribution; ii) the resulting estimators may have large variance (Example: estimating a very small probability).

4.5 Importance Sampling

The simplest form of importance sampling is based on the identity

$$\int f(\theta)p(\theta)d\theta = \int f(\theta)p(\theta)/q(\theta)q(\theta)d\theta.$$

Here q, known as the "importance sampling distribution" can be any distribution whose support contains the support of p (i.e. q is non-zero wherever p is non-zero).

This formulation leads to

$$\int f(\theta)p(\theta)d\theta \approx (1/M)\sum_{i=1}^{M} f(\theta^{(i)})p(\theta^{(i)})/q(\theta^{(i)})$$

where $\theta^{(1)}, \ldots, \theta^{(M)} \sim q(\theta)$. If q is well chosen then this estimator can be very much less variable than an estimator based on standard Monte Carlo. (But if q is badly chosen it can be worse, and even have infinite variance!)

Another form of importance sampling can be useful if p can be computed only up to a constant of proportionality (e.g. if p is the posterior, and the prior and likelihood are known, but the normalizing constant is tricky). This form is as follows. If $\theta^{(1)}, \ldots, \theta^{(M)} \sim q(\theta)$ then the distribution with weight $w^{(i)} \propto p(\theta)p(x|\theta)/q(\theta)$ on $\theta^{(i)}$, with the weights normalised to sum to 1, can be thought of as an approximation to the posterior distribution for θ. [Here the distribution q is known as the Importance Sampling Distribution, and the weights are known as Importance Weights.]

In particular

$$E(f(\theta)|x) \approx \sum_{m} w^{(m)} f(\theta^{(m)}). \tag{1}$$

Note the special cases where q is the posterior distribution (leading to uniform weights), and also where q is the prior (leading to weights proportional to the likelihood).

Choice of q

One would like to choose q so that the variance of the estimator (**??**) is small. However, this variance depends on the function f, and we might want to choose one q for many different fs. Typically one tries to choose q so that it approximates the posterior, which leads to approximately uniform weights. A rule of thumb is that q should have "longer tails" than the posterior: you don't want very large weights that have very small probability under q. This can be difficult to achieve, particularly if θ is high-dimensional. One trick is to make q a mixture of two distributions: a long-tailed distribution (e.g. the prior) and another distribution, which attempts to approximate the posterior.

4.6 Markov chain Monte Carlo (MCMC)

MCMC is a very widely-used technique for simulating samples from the posterior distribution.

The basic idea is to simulate a Markov chain, whose stationary distribution is the posterior distribution. By simulating the Markov chain for long enough, one obtains samples that are "approximately" from the posterior distribution. The question of how long is long enough is generally difficult to answer, and is not one we will spend much time addressing here.

Metropolis–Hastings Algorithm

Let $q(\theta \to \theta')$ denote a transition density, which is a way of generating new states θ' given the current state θ. Let π denote the target distribution, up to a constant of proportionality (e.g. to simulate from the posterior, $\pi = p(\theta)p(x|\theta)$).

Now consider the following Markov chain:

1. Start with an initial value $\theta^{(0)}$; set $t = 0$.
2. Given the current value of $\theta^{(t)} = \theta$, generate a proposed new value θ' according to $q(\theta \to \cdot)$.
3. Define the acceptance probability A by
$$A = \min\left(1, \frac{\pi(\theta')q(\theta' \to \theta)}{\pi(\theta)q(\theta \to \theta')}\right).$$
4. With probability A set $\theta^{(t+1)} = \theta'$; otherwise set $\theta^{(t+1)} = \theta$.
5. Increase t; return to step 2.

Note: q is referred to as the *proposal distribution*. The probability A is the *acceptance probability*.

If q is chosen to be irreducible and aperiodic, then this Markov chain has stationary distribution $\pi()$. Proof: check detailed balance condition, $\pi(\theta)p(\theta \to \theta') = \pi(\theta')p(\theta' \to \theta)$.

Ideally you want q to have two properties

- It proposes large moves.
- It leads to high average acceptance probabilities, A.

In general these two properties are competing: it is difficult to achieve both. As a result, one generally tries to choose q so that it gives "intermediate" average values for A (e.g. a few percent in typical problems?). [An exception to this is Gibbs sampling, below, where $A = 1$.]

Example: Random Walk Metropolis-Hastings

Perhaps the most common type of MH sampler is the "random walk" MH sampler, where the proposal distribution q involves adding a symmetric random number (e.g. $U[-\epsilon, \epsilon]$ or $N(0, \epsilon^2)$) to the current value of θ. In this case, the terms involving q in the acceptance probability cancel, due to symmetry. The value of ϵ determines the typical size of the proposed move, and hence the typical value for A.

Component-wise Metropolis

The MH algorithm works as is for θ a multi-dimensional parameter. However, for multi-dimensional parameters, an alternative is to update each component of the vector, one at a time.

For example, if $\theta = (\theta_1, \theta_2)$ then it is OK to use a MH update step for θ_1, keeping the current value of θ_2 fixed, and then to use an MH step for θ_2, keeping θ_1 fixed.

The advantage of component-wise updates is that it can make it easier to find proposal distributions q that lead to moderate values of the acceptance probability A (not too big or too small).

Gibbs Sampling

Suppose that $\theta = (\theta_1, \ldots, \theta_k)$ is a vector of unknown parameters. Then consider the following Markov chain

1. Start with an initial value $\theta^{(0)}$; set $t = 0$.

2. Sample $\theta_1^{(t+1)}$ from $p(\theta_1|x, \theta_2^{(t)}, \ldots, \theta_k^{(t)})$.

3. Sample $\theta_2^{(t+1)}$ from $p(\theta_2|x, \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)})$.

4. ...

5. Sample $\theta_k^{(t+1)}$ from $p(\theta_k|x, \theta_1^{(t+1)}, \ldots, \theta_{k-1}^{(t+1)})$.

6. Increase t and return to 2.

Note that at each step a component of the unknown parameter is sampled from its full conditional distribution, given the data, and the current value of all other components of the parameter.

It is easy to show that this Markov chain has stationary distribution $p(\theta|x)$. [Proof?]

Indeed, this Markov chain is a special case of component-wise MH sampling: using the full conditional distributions as the proposal distribution in an MH sampler gives acceptance probability $A = 1$ [Exercise].