Bayesian Inference I

Loukia Meligkotsidou,
National and Kapodistrian University of Athens

MSc in Biostatistics,
Department of Mathematics and School of Medicine

Outline of the course

This course provides theory and practice of the Bayesian approach to statistical inference. Applications are performed with the statistical package R.

Topics:

- Bayesian Updating through Bayes' Theorem
- Prior Distributions
- Multi-parameter Problems
- Summarizing Posterior Information
- Prediction
- The Gibbs Sampler

Introduction

Commonly the purpose of formulating a statistical model is to make *predictions* about future values of the process.

In making predictions about future values on the basis of an estimated model there are two sources of uncertainty:

- Uncertainty in the parameter values which have been estimated from past data; and
- Uncertainty due to the fact that any future value is itself a **random event**.

In classical statistics it is usual to fit a model to the past data, and then make predictions of future values on the assumption that this model is correct (*estimative* approach). Only the second source of uncertainty is included in the analysis, leading to estimates which are believed to be more precise than they really are.

The Predictive Density

Within Bayesian inference it is straightforward to allow for both sources of uncertainty by simply averaging over the uncertainty in the parameter estimates, the information of which is completely contained in the posterior distribution.

So, suppose we have past observations $x = (x_1, \ldots, x_n)$ of a variable with density function (or likelihood) $f(x|\theta)$ and we wish to make inferences about the distribution of a future value of a random variable Y from this same model.

With a prior distribution $f(\theta)$, Bayes' theorem leads to a posterior distribution $f(\theta|x)$. Then the *predictive density* of y given x is:

$$f(y|x) = \int f(y|\theta)f(\theta|x)d\theta = E\left[f(y|\theta)|x\right].$$

Thus the predictive density is the integral (expectation) of the likelihood of y with respect to the posterior.

Derivation of the Predictive Density

Note: the derivation of the predictive distribution is simply based on the usual laws of probability manipulation, and has a straightforward interpretation itself in terms of probabilities.

The r.v. Y need not come from the same distribution as the observations x. It is important however that, given θ, we assume that Y is independent of x. Therefore,

Joint density of y and x, given θ: $f(y, x|\theta) = f(y|\theta)f(x|\theta)$,

Derivation of the Predictive Density

Note: the derivation of the predictive distribution is simply based on the usual laws of probability manipulation, and has a straightforward interpretation itself in terms of probabilities.

The r.v. Y need not come from the same distribution as the observations x. It is important however that, given θ, we assume that Y is independent of x. Therefore,

Joint density of y and x, given θ: $f(y, x|\theta) = f(y|\theta)f(x|\theta)$,
Joint density of y, x and θ: $f(y, x, \theta) = f(y|\theta)f(x|\theta)f(\theta)$.

Derivation of the Predictive Density

Note: the derivation of the predictive distribution is simply based on the usual laws of probability manipulation, and has a straightforward interpretation itself in terms of probabilities.

The r.v. Y need not come from the same distribution as the observations x. It is important however that, given θ, we assume that Y is independent of x. Therefore,

Joint density of y and x, given θ: $f(y, x|\theta) = f(y|\theta)f(x|\theta)$,
Joint density of y, x and θ: $f(y, x, \theta) = f(y|\theta)f(x|\theta)f(\theta)$.
Then

$$f(y, \theta|x) = \frac{f(y|\theta)f(x|\theta)f(\theta)}{f(x)} = f(y|\theta)f(\theta|x),$$

Derivation of the Predictive Density

Note: the derivation of the predictive distribution is simply based on the usual laws of probability manipulation, and has a straightforward interpretation itself in terms of probabilities.

The r.v. Y need not come from the same distribution as the observations x. It is important however that, given θ, we assume that Y is independent of x. Therefore,

Joint density of y and x, given θ: $f(y, x|\theta) = f(y|\theta)f(x|\theta)$,
Joint density of y, x and θ: $f(y, x, \theta) = f(y|\theta)f(x|\theta)f(\theta)$.
Then

$$f(y, \theta|x) = \frac{f(y|\theta)f(x|\theta)f(\theta)}{f(x)} = f(y|\theta)f(\theta|x),$$

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

Classical Approach

The corresponding approach in classical statistics would be to obtain the maximum likelihood estimate $\hat{\theta}$ of θ and to base inference on the distribution $f(y|\hat{\theta})$, the *estimative* distribution.

This makes no allowance for the variability incurred as a result of estimating θ, and so gives a false sense of precision (the predictive density $f(y|x)$ is more variable by averaging across the posterior distribution for θ).

Classical Approach

The corresponding approach in classical statistics would be to obtain the maximum likelihood estimate $\hat{\theta}$ of θ and to base inference on the distribution $f(y|\hat{\theta})$, the *estimative* distribution.

This makes no allowance for the variability incurred as a result of estimating θ, and so gives a false sense of precision (the predictive density $f(y|x)$ is more variable by averaging across the posterior distribution for θ).

Note: You CANNOT remove a constant of proportionality in $f(y|\theta)$, while it is usually simplest to use the (normalised) posterior distribution $f(\theta|x)$ in

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

(If you use the posterior up to a constant of proportionality, then you will also get $f(y|x)$ up to a constant of proportionality).

Example. Binomial Sample

Suppose we have made an observation $x \sim Binomial(n, \theta)$ and our (conjugate) prior for θ is $\theta \sim Beta(p, q)$. Then, we have shown, the posterior for θ is given by:

$$\theta | x \sim Beta(p + x, q + n - x).$$

Now, suppose we intend to make a further N observations in the future, and we let z be the number of successes in those N trials, so that $z|\theta \sim Binomial(N, \theta)$. So, we have the likelihood for our future observation:

$$f(z|\theta) = \binom{N}{z} \theta^z (1 - \theta)^{N-z}.$$

The Predictive Distribution

For $z = 0, 1, \ldots, N$,

$$f(z|x) = \int_0^1 \binom{N}{z} \theta^z (1-\theta)^{N-z} \times \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{B(p+x, q+n-x)} d\theta$$

The Predictive Distribution

For $z = 0, 1, \ldots, N$,

$$\begin{aligned} f(z|x) &= \int_0^1 \binom{N}{z} \theta^z (1-\theta)^{N-z} \times \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{B(p+x, q+n-x)} d\theta \\ &= \binom{N}{z} \frac{1}{B(P,Q)} \int_0^1 \theta^{P+z-1}(1-\theta)^{Q+N-z-1} \end{aligned}$$

The Predictive Distribution

For $z = 0, 1, \ldots, N$,

$$f(z|x) = \int_0^1 \binom{N}{z} \theta^z (1-\theta)^{N-z} \times \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{B(p+x, q+n-x)} d\theta$$

$$= \binom{N}{z} \frac{1}{B(P,Q)} \int_0^1 \theta^{P+z-1}(1-\theta)^{Q+N-z-1}$$

$$= \binom{N}{z} \frac{B(P+z, Q+N-z)}{B(P,Q)}.$$

This is, in fact, known as a Beta–binomial distribution.

Example. Gamma Sample

Suppose $X_1, \ldots X_n$ are independent variables having the $Gamma(k, \theta)$ distribution, where k is known, and we use the conjugate prior $\theta \sim Gamma(p, q)$:

$$f(\theta) \propto \theta^{p-1} \exp\{-q\theta\}$$

leading via Bayes' theorem to
$\theta|x \sim Gamma(p + nk, q + \Sigma x_i) = Gamma(G, H)$.

The likelihood for a future observation y is

$$f(y|\theta) = \frac{\theta^k y^{k-1} \exp\{-\theta y\}}{\Gamma(k)}$$

The Predictive Distribution

$$f(y|x) = \int_0^\infty \frac{\theta^k y^{k-1} \exp\{-\theta y\}}{\Gamma(k)} \times \frac{H^G \theta^{G-1} \exp\{-H\theta\}}{\Gamma(G)} d\theta$$

The Predictive Distribution

$$f(y|x) = \int_0^\infty \frac{\theta^k y^{k-1} \exp\{-\theta y\}}{\Gamma(k)} \times \frac{H^G \theta^{G-1} \exp\{-H\theta\}}{\Gamma(G)} d\theta$$

$$= \frac{H^G y^{k-1}}{\Gamma(k)\Gamma(G)} \int_0^\infty \theta^{k+G-1} \exp\{-\theta(y+H)\} d\theta$$

The Predictive Distribution

$$
\begin{aligned}
f(y|x) &= \int_0^\infty \frac{\theta^k y^{k-1} \exp\{-\theta y\}}{\Gamma(k)} \times \frac{H^G \theta^{G-1} \exp\{-H\theta\}}{\Gamma(G)} d\theta \\
&= \frac{H^G y^{k-1}}{\Gamma(k)\Gamma(G)} \int_0^\infty \theta^{k+G-1} \exp\{-\theta(y+H)\} d\theta \\
&= \frac{H^G y^{k-1}}{\Gamma(k)\Gamma(G)} \frac{\Gamma(k+G)}{(y+H)^{k+G}} = \frac{H^G y^{k-1}}{B(k,G)(H+y)^{G+k}}, \quad y > 0.
\end{aligned}
$$

The Predictive Distribution

$$
\begin{aligned}
f(y|x) &= \int_0^\infty \frac{\theta^k y^{k-1} \exp\{-\theta y\}}{\Gamma(k)} \times \frac{H^G \theta^{G-1} \exp\{-H\theta\}}{\Gamma(G)} d\theta \\
&= \frac{H^G y^{k-1}}{\Gamma(k)\Gamma(G)} \int_0^\infty \theta^{k+G-1} \exp\{-\theta(y+H)\} d\theta \\
&= \frac{H^G y^{k-1}}{\Gamma(k)\Gamma(G)} \frac{\Gamma(k+G)}{(y+H)^{k+G}} = \frac{H^G y^{k-1}}{B(k,G)(H+y)^{G+k}}, \quad y > 0.
\end{aligned}
$$

We can relate $f(y \mid x)$ to a standard distribution by writing

$$
Y = (H\nu_1/\nu_2) F_{\nu 1, \nu 2},
$$

where $\nu_1 = 2k$ and $\nu_2 = 2G$ and $F_{\nu 1, \nu 2}$ has the Fisher ‘F’ distribution.