Bayesian Inference

Loukia Meligkotsidou,
National and Kapodistrian University of Athens

MSc in Statistics and Operational Research,
Department of Mathematics

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
- The Multivariate Normal Model
- Prediction
- Asymptotics
- Markov chain Monte Carlo Methods

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 $x = (x_1, \dots, x_n)$ 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 $z = 0, 1, \ldots, N$,

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

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

The Predictive Distribution

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

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

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

Example. Gamma Sample

Suppose $X_1, \dots X_n$ 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

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

The Predictive Distribution

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

The Predictive Distribution

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

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.

Exercise

Consider an event that occurred n times in n trials. Laplace suggested that the probability of the event happening in the next trial is $\frac{n+1}{n+2}$. Confirm this suggestion assuming a Uniform(0,1) prior distribution for the probability of the event.

Exercise

Consider an event that occurred n times in n trials. Laplace suggested that the probability of the event happening in the next trial is $\frac{n+1}{n+2}$. Confirm this suggestion assuming a Uniform(0,1) prior distribution for the probability of the event.

Solution. Let X be the number of occurrences of the event in n trials and *theta* be the probability of the event. We have

$$X \sim Binomial(n, \theta).$$

We have observed $x = n$.

Exercise

Consider an event that occurred n times in n trials. Laplace suggested that the probability of the event happening in the next trial is $\frac{n+1}{n+2}$. Confirm this suggestion assuming a Uniform(0,1) prior distribution for the probability of the event.

Solution. Let X be the number of occurrences of the event in n trials and $theta$ be the probability of the event. We have

$$X \sim Binomial(n, \theta).$$

We have observed $x = n$.

prior: $\theta \sim Beta(1, 1)$
posterior: $\theta \mid x \sim Beta(1 + x, 1 + n - x)$

Exercise

Consider an event that occurred n times in n trials. Laplace suggested that the probability of the event happening in the next trial is $\frac{n+1}{n+2}$. Confirm this suggestion assuming a Uniform(0,1) prior distribution for the probability of the event.

Solution. Let X be the number of occurrences of the event in n trials and *theta* be the probability of the event. We have

$$X \sim Binomial(n, \theta).$$

We have observed $x = n$.

prior: $\theta \sim Beta(1, 1)$
posterior: $\theta \mid x \sim Beta(1 + x, 1 + n - x)$
$\theta \mid x = n \sim Beta(1 + n, 1)$

Exercise

Likelihood of the future observation y (Bernoulli):

$$f(y \mid \theta) = \theta^y (1-\theta)^{1-y}, \quad y = 0, 1$$

Predictive distribution:

$$f(y \mid x = n) \quad = \quad \int_0^1 f(y \mid \theta) f(\theta \mid x = n) d\theta$$

Exercise

Likelihood of the future observation y (Bernoulli):

$$f(y \mid \theta) = \theta^y (1-\theta)^{1-y}, \quad y = 0, 1$$

Predictive distribution:

$$\begin{aligned} f(y \mid x = n) &= \int_0^1 f(y \mid \theta) f(\theta \mid x = n) d\theta \\ &= \int_0^1 \theta^y (1-\theta)^{1-y} \frac{\theta^n}{B(n+1,1)} d\theta \end{aligned}$$

Exercise

Likelihood of the future observation y (Bernoulli):

$$f(y \mid \theta) = \theta^y (1-\theta)^{1-y}, \quad y = 0, 1$$

Predictive distribution:

$$\begin{aligned} f(y \mid x = n) &= \int_0^1 f(y \mid \theta) f(\theta \mid x = n) d\theta \\ &= \int_0^1 \theta^y (1-\theta)^{1-y} \frac{\theta^n}{B(n+1,1)} d\theta \\ &= \frac{1}{B(n+1,1)} \int_0^1 \theta^{n+y} (1-\theta)^{1-y} d\theta \end{aligned}$$

Exercise

Likelihood of the future observation y (Bernoulli):

$$f(y \mid \theta) = \theta^y (1-\theta)^{1-y}, \quad y = 0, 1$$

Predictive distribution:

$$\begin{aligned} f(y \mid x = n) &= \int_0^1 f(y \mid \theta) f(\theta \mid x = n) d\theta \\ &= \int_0^1 \theta^y (1-\theta)^{1-y} \frac{\theta^n}{B(n+1,1)} d\theta \\ &= \frac{1}{B(n+1,1)} \int_0^1 \theta^{n+y} (1-\theta)^{1-y} d\theta \\ &= \frac{B(n+y+1, 1-y+1)}{B(n+1,1)}, \quad y = 0, 1. \end{aligned}$$

Exercise

Probability of the event happening in the next trial:

$$\Pr(Y = 1 \mid x = n) \quad = \quad f(y = 1 \mid x = n)$$

Exercise

Probability of the event happening in the next trial:

$$\begin{aligned}
\Pr(Y=1 \mid x=n) &= f(y=1 \mid x=n) \\
&= \frac{B(n+2,1)}{B(n+1,1)} = \frac{\frac{\Gamma(n+2)\Gamma(1)}{\Gamma(n+3)}}{\frac{\Gamma(n+1)\Gamma(1)}{\Gamma(n+2)}} \\
&= \frac{\Gamma(n+2)\Gamma(n+2)}{\Gamma(n+3)\Gamma(n+1)}
\end{aligned}$$

Exercise

Probability of the event happening in the next trial:

$$
\begin{aligned}
\Pr(Y=1 \mid x=n) &= f(y=1 \mid x=n) \\
&= \frac{B(n+2,1)}{B(n+1,1)} = \frac{\frac{\Gamma(n+2)\Gamma(1)}{\Gamma(n+3)}}{\frac{\Gamma(n+1)\Gamma(1)}{\Gamma(n+2)}} \\
&= \frac{\Gamma(n+2)\Gamma(n+2)}{\Gamma(n+3)\Gamma(n+1)} \\
&= \frac{(n+1)!(n+1)!}{(n+2)!n!} = \frac{n+1}{n+2}.
\end{aligned}
$$

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
- The Multivariate Normal Model
- Prediction
- Asymptotics
- Markov chain Monte Carlo Methods

Asymptotics

Looking back at the conjugate analysis for the Normal mean θ with $X_1 \ldots X_n \sim N(\theta, \tau^{-1})$, we obtain for a prior $\theta \sim N(b, c^{-1})$

$$\theta | x \sim N(\frac{cb + n\tau\overline{x}}{c + n\tau}, \frac{1}{c + n\tau})$$

Now, as n becomes large, this becomes:

$$\theta | x \sim N(\overline{x}, \frac{1}{n\tau}) = N(\overline{x}, \frac{\sigma^2}{n})$$

Thus, the effect of the prior disappears, and the posterior is determined solely by the data. Moreover, the posterior distribution becomes increasingly more concentrated around $\overline{x}$, which by the strong law of large numbers converges to the true value of θ.

Consistency

If the true value of θ is θ_0, and the prior probability of θ_0 (or in the continuous case an arbitrary neighbourhood of θ_0) is not zero, then with increasing amounts of data x, the posterior probability that $\theta = \theta_0$ (or lies in a neighbourhood of θ_0) tends to unity.

Consistency

If the true value of θ is θ_0, and the prior probability of θ_0 (or in the continuous case an arbitrary neighbourhood of θ_0) is not zero, then with increasing amounts of data x, the posterior probability that $\theta = \theta_0$ (or lies in a neighbourhood of θ_0) tends to unity.

Proof. Let $x_1, \ldots x_n$ be IID observations, each with distribution $g(x|\theta)$. Then the posterior density is

$$f(\theta|x) \quad \propto \quad f(\theta) \prod_{i=1}^{n} g(x_i|\theta) = f(\theta) \exp\left\{\sum_{i=1}^{n} \log g(x_i|\theta)\right\}$$

Consistency

If the true value of θ is θ_0, and the prior probability of θ_0 (or in the continuous case an arbitrary neighbourhood of θ_0) is not zero, then with increasing amounts of data x, the posterior probability that $\theta = \theta_0$ (or lies in a neighbourhood of θ_0) tends to unity.

Proof. Let $x_1, \dots x_n$ be IID observations, each with distribution $g(x|\theta)$. Then the posterior density is

$$
\begin{aligned}
f(\theta|x) &\propto f(\theta)\prod_{i=1}^{n} g(x_i|\theta) = f(\theta)\exp\left\{\sum_{i=1}^{n}\log g(x_i|\theta)\right\} \\
&= f(\theta)\exp\{n\bar{\ell}_n(\theta)\}, \bar{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log g(x_i|\theta) \\
&\propto f(\theta)\exp\{n[\bar{\ell}_n(\theta) - \bar{\ell}_n(\theta_0)]\}.
\end{aligned}
$$

For fixed θ, let $h(x_i) = \log g(x_i \mid \theta) - \log g(x_i \mid \theta_0)$ (function of x_i). Then, $\overline{\ell}_n(\theta) - \overline{\ell}_n(\theta_0)$ is the sample mean of $h(x_i)$, $i = 1, \ldots, n$, and so converges in probability to the expectation of $h(x)$:

$$
\begin{aligned}
E\left[h(x)\right] &= \int h(x) g(x \mid \theta_0) dx \\
&= \int \{\log g(x|\theta) - \log g(x|\theta_0)\} g(x|\theta_0) dx.
\end{aligned}
$$

It can be shown that $E\left[h(x)\right] \leq 0$ and $E\left[h(x)\right] = 0$ for $\theta = \theta_0$. Thus, for $\theta \neq \theta_0$, it follows that $\exp\{n[\overline{\ell}_n(\theta) - \overline{\ell}_n(\theta_0)]\}$ tends to 0 with probability 1 as $n \to \infty$, but remains at 1 for $\theta = \theta_0$.

Thus, as long as $f(\theta_0) \neq 0$, the posterior
$f(\theta \mid x) \propto f(\theta) \exp\{n[\overline{\ell}_n(\theta) - \overline{\ell}_n(\theta_0)]\} \to 1$, for $\theta = \theta_0$ and is 0 everywhere else.
(As long as the prior gives non–zero weight to the true value of θ, eventually, the posterior will concentrate on the true value).

Asymptotic Normality

For continuous θ, as n increases $\exp\{n[\bar{\ell}_n(\theta) - \bar{\ell}_n(\theta_0)]\}$ is negligibly small on all but a vanishingly small neighbourhood of θ_0.

Hence $f(\theta)$ can be regarded as constant over this neighbourhood and we obtain

$$\begin{aligned} f(\theta|x_1, \ldots x_n) &\propto \exp\{n\bar{\ell}_n(\theta)\} \\ &\propto \exp\{n[\bar{\ell}_n(\theta) - \bar{\ell}_n(\hat{\theta})]\}, \end{aligned}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ.

Now we expand the exponent in this expression as a Taylor series about $\hat{\theta}$:

$$\begin{aligned} n[\bar{\ell}_n(\theta) - \bar{\ell}_n(\hat{\theta})] &= n[\bar{\ell}_n(\hat{\theta}) - \bar{\ell}_n(\hat{\theta})] + n(\theta - \hat{\theta})\bar{\ell}'_n(\hat{\theta}) + \frac{n}{2!}(\theta - \hat{\theta})^2\, \bar{\ell}''(\hat{\theta}) + \ldots \\ &\approx \frac{n}{2}(\theta - \hat{\theta})^2 \bar{\ell}''(\hat{\theta}). \end{aligned}$$

Then

$$f(\theta|x) \quad \propto \quad \exp\{-\frac{n}{2}\left[-\bar{\ell}''(\hat{\theta})\right] \times (\theta - \hat{\theta})^2\}.$$

That is

$$\theta|x \sim N(\hat{\theta}, \left[-n\bar{\ell}''(\hat{\theta})\right]^{-1}).$$

So, as $n \to \infty$, the distribution of the posterior is approximately Normal about the m.l.e. $\hat{\theta}$, and with variance given by minus the second derivative of the log likelihood at the maximum. Notice again, that this result is true independently of the prior specification, provided the prior is not zero at the true value.

Example. Normal Mean

Let $X_1, \dots, X_n$ be a set of independent variables from $N(\theta, \sigma^2)$, where σ^2 is known. This gives the likelihood

$$f(x|\theta) \propto \exp\{-\frac{\sum_{i=1}^{n}(x_i - \theta)^2}{2\sigma^2}\}$$

Thus, we can take

$$\log(f(x|\theta)) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2 + c$$

so that

$$\frac{d\log(f(x|\theta))}{d\theta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta)$$

and

$$\frac{d^2\log(f(x|\theta))}{d^2\theta} = -n/\sigma^2.$$

Example. Normal Mean

Consequently, the m.l.e. $\hat{\theta} = \overline{x}$ and $-n\bar{\ell}''(\hat{\theta}) = n/\sigma^2$. Hence, asymptotically as $n \to \infty$,

$$\theta | x \sim N(\overline{x}, \sigma^2/n).$$

This is true for *any* prior distribution which places non–zero probability around the true value of θ.

Example. Binomial Sample

Consider the likelihood model $x \sim Binomial(n, \theta)$. So

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}; \qquad x = 0, \ldots, n.$$

Thus

$$\log(f(x|\theta)) = x \log\theta + (n-x)\log(1-\theta).$$

So,

$$\frac{d \log(f(x|\theta))}{d\theta} = \frac{x}{\theta} - \frac{(n-x)}{1-\theta}$$

and

$$\frac{d^2 \log \ell(\theta)}{d^2\theta} = \frac{-x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2}.$$

Example. Binomial Sample

Consequently, $\hat{\theta} = x/n$ and

$$-n\bar{\ell}''(\hat{\theta}) = \frac{n\hat{\theta}}{\hat{\theta}^2} - \frac{n(1-\hat{\theta})}{(1-\hat{\theta})^2} = \frac{n}{\hat{\theta}(1-\hat{\theta})}.$$

Thus, as $n \to \infty$,

$$\theta | x \sim N(\frac{x}{n}, \frac{\frac{x}{n}(1-\frac{x}{n})}{n}).$$