Introduction

Goal Make inference about the parameters and the structure of a statistical model using data, and quantify all relevant uncertainty with probabilities.

Step 1 Define a probability model for all components of the statistical model, observables or not.

Step 2 Conditional on the observed data, produce a conditional posterior distribution for the unobserved quantities: parameters, missing values, future values, latent variables or any other. Every unobservable quantity is treated as a random variable.

Step 3 Evaluate the goodness of fit of the model.

Probabilistic Model: Denote θ the unobserved variables and y the observable ones. We build a joint distribution for θ and y as

$$p(y, \theta) = p(\theta)p(y|\theta)$$

$p(\theta)$ is the **prior** or **initial** distribution for θ and $p(y|\theta)$ is the distribution of the observables given the parameters or **likelihood**.

Using **Bayes Rule** we obtain the **posterior** distribution of θ given y as

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

$p(y)$ is the **marginal** for y and is given as

$$p(y) = \int_{\Theta} p(\theta)p(y|\theta)d\theta$$

Given y, $p(y)$ is just a constant, so it is usually disregarded when the goal is to explore the posterior distribution.

Posterior Inference

Bayesian inference is based on the $p(\theta|y)$. Generally speaking we need to calculate functionals of the type

$$\int_{\Theta} h(\theta)p(\theta|y)d\theta$$

for a function h. Letting $h(\theta) = \theta$ gives the posterior mean. Letting $h(\theta) = \mathbb{1}_A(\theta)$ gives the posterior probability of the set A.

Marginals

When θ is a multivariate parameter we often are interested in only some of its components. Let $\theta = (\theta_1, \theta_2)$, then, the posterior distribution of θ_1 is given as

$$p(\theta_1|y) = \int_{\Theta_2} p(\theta_1, \theta_2|y)d\theta_2$$

Predictions

Let z be a, yet, unobserved variable whose value we need to predict. Predictions are based on the **posterior predictive** distribution of z given y.

$$p(z|y) = \int_\Theta p(z, \theta|y) d\theta = \int_\Theta p(z|\theta, y) p(\theta|y) d\theta$$

In many cases this formula is simplified by the fact that $p(z|\theta, y) = p(z|\theta)$.

Conditional Independence $y_1, \ldots, y_n$ correspond to the tosses of a coin whose probability of landing tails is θ. Then

$$p(y_{n+1}|\theta, y_1, \ldots, y_n) = p(y_{n+1}|\theta) = \theta$$

Conditional Dependence Consider an AR(1), $x_{n+1} = \alpha x_n + \varepsilon_{n+1},\ \varepsilon_{n+1} \sim N(0, \sigma^2)$. Then

$$p(x_{n+1}|\alpha, \sigma^2, x_1, \ldots, x_n) = p(x_{n+1}|\alpha, \sigma^2, x_n) = N(\alpha x_n, \sigma^2)$$

Likelihood Principle

When calculating the posterior distribution of θ we can multiply the likelihood by a constant and the resulting posterior will be unchanged.

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)K}{\int_\Theta p(\theta)p(y|\theta)K\,d\theta}$$

K cancels out. Thus, inferences that use the posterior distribution satisfy the **Likelihood Principle**, according to which inference should not change when the likelihood is multiplied by a constant.

An example of the likelihood principle in operation is making inference about the probability of success in a binary trial. If we stop the process after a fixed number of trials we use a Binomial likelihood. If we stop the process after observing a fixed number of successes then we use a Negative Binomial. In a Bayesian setting both likelihoods will produce the same results, since they are proportional.

Exchangeability

Suppose that, in thinking of the joint distribution $p(x_1, \ldots, x_n)$, we conclude that the subscripts of the random variables are uninformative. This motivates the idea of exchangeability.

Definition: The random variables $X_1, \ldots X_n$ are exchangeable if

$$p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)}),$$
$$\pi \quad \text{any permutation on} \quad \{1, \ldots, n\}$$

One consequence of this definition is that all marginal densities are the same. Clearly, i.i.d. implies exchangeability.

Exchangeability is usually modeled using conditional independence. This is based on the following theorem.

Theorem (de Finetti's): If $X_1, X_2, \ldots$ is an infinitely exchangeable sequence of 0-1 random quantities, then there is a probability distribution F such that

$$p(x_1, \ldots, x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} dF(\theta)$$

The theorem implies that any exchangeable sequence of binary random variables can be represented as a sequence of independent Bernoulli variables, conditional on θ, where θ follows a probability distribution F.

The representation theorem can be extended to the more general setting of real valued random variables as

$$p(x_1, \ldots, x_n) = \int_\Theta \prod_{i=1}^{n} p(x_i|\theta) dF(\theta)$$

De Finetti's representation provides a formal justification to the use of prior distributions on θ.

Conditional independence provides exchangeability and is a powerful tool for model building. It is at the basis of most hierarchical models.

Conjugate Families

One of the most convenient ways of specifying a parametric prior is to use a conjugate family.

Definition: Let $\mathcal{F} = \{p(x|\theta), \theta \in \Theta\}$ be a family of sampling distributions. A class $\mathcal{P}$ is said to be a conjugate family for $\mathcal{F}$ if for all $p \in \mathcal{F}$ and $p(\theta) \in \mathcal{P}$ then $p(\theta|x) \in \mathcal{P}$.

So prior and posterior distributions belong to the same parametric family.

Example: Consider a sample $Y \sim Bin(n, \theta)$, then the conjugate prior for θ is a Beta distribution, in fact

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad p(Y|\theta) \propto \theta^{Y}(1-\theta)^{n-Y}$$

and

$$p(\theta|Y) \propto \theta^{Y+\alpha-1}(1-\theta)^{n-Y+\beta-1}$$

so that, we start with a $Beta(\alpha, \beta)$ and update it to a $Beta(Y+\alpha, n-Y+\beta)$.

In general, inference with conjugate distributions consist of updating the parameters of the prior once the data become available.

Example: Consider the exponential family given by

$$p(x|\theta) = a(x) \exp \left\{ \sum_{j=1}^{r} U_j(x)\phi_j(\theta) + b(\theta) \right\}$$

then, the conjugate family is given by

$$p(\theta) = k(\alpha, \beta) \exp \left\{ \sum_{j=1}^{r} \alpha_j \phi_j(\theta) + \beta b(\theta) \right\}$$

leading to the posterior

$$p(\theta|x) = k^*(\alpha, \beta) \exp \left\{ \sum_{j=1}^{r} (U_j(x) + \alpha_j)\phi_j(\theta) + (\beta + 1)b(\theta) \right\}$$

Example: Consider a random sample of size n from a normal distribution with known variance. Then the likelihood is proportional to

$$\exp\left\{-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right\}$$

The conjugate prior is given by a normal, say, $p(\theta) = N(\theta|\mu, \tau^2)$. Then the posterior distribution is also a normal $N(\theta|\mu_*, \tau_*^2)$ where

$$\mu_* = \frac{n\sigma^{-2}\overline{x} + \tau^{-2}\mu}{n\sigma^{-2} + \tau^{-2}} \qquad \tau_*^{-2} = n\sigma^{-2} + \tau^{-2}$$

The posterior mean is a convex combination of μ and $\overline{x}$

$$\mu_* = t\overline{x} + (1-t)\mu, \quad t = \frac{n\sigma^{-2}}{n\sigma^{-2} + \tau^{-2}}, \quad t \in (0,1)$$

Example: When the variance of the normal likelihood is also unknown we obtain a bivariate prior by conditioning. Let $\phi = \sigma^{-2}$, then the prior $p(\mu, \phi) = p(\mu|\phi)p(\phi)$ is proportional to

$$\phi^{1/2} \exp\left\{-\frac{c_0\phi}{2}(\theta - \mu_0)^2\right\} \phi^{n_0/2-1} \exp\left\{-\frac{n_0\sigma_0^2}{2}\phi\right\}$$

the product of a $N(\theta|\mu_0, 1/c_0\phi)$ and $Gam(\phi|n_0/2, n_0\sigma_0^2/2)$. Integrating ϕ, we have that the marginal distribution of θ is a $t_{n_0}(\theta|\mu_0, \sigma_0^2/c_0)$. A posteriori the parameters are updated to

$$\mu_1 = \frac{c_0\mu + n\bar{x}}{c_0+n}, \quad c_1 = c_0 + n, \quad n_1 = n_0 + n,$$
$$n_1\sigma_1^2 = n_0\sigma_0^2 + ns^2 + \frac{c_0 n}{c_0+n}(\mu_0 - \bar{x})^2$$

preserving the Normal-Gamma structure.

Jefferey's Prior

Definition: For an observation X with probability density function $p(x|\theta)$ the Jeffreys' non-informative prior is given by

$$p(\theta) \propto |I(\theta)|^{1/2}$$

where $I(\theta)$ denotes the Fisher Information Matrix which has components

$$I_{ij}(\theta) = E\left[-\frac{\partial^2 \log p(X|\theta)}{\partial\theta_i \partial\theta_j}\right]$$

Model Comparison

Suppose we want to compare two models

$$M_1 : p_1(y|\theta_1) \quad \text{and} \quad p_1(\theta_1)$$

$$M_2 : p_2(y|\theta_2) \quad \text{and} \quad p_2(\theta_2)$$

then we assume that prior probabilities for each model $p(M_1)$ and $p(M_2)$ are available and update them to obtain

$$p(M_1|y) = p(y|M_1)p(M_1) \quad p(M_2|y) = p(y|M_2)p(M_2)$$

where

$$p(y|M_i) = \int_{\Theta_i} p_i(\theta_i|M_i)p_i(y|\theta_i)d\theta_i$$

The ratio of the two posterior probabilities is

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(M_1)}{p(M_2)}\frac{p(y|M_1)}{p(y|M_2)} = \frac{p(M_1)}{p(M_2)}B_{12}(y)$$

Thus, the odds ratio is updated by the **Bayes Factor** $B_{12}(y)$.

Simulations

A very natural way of making inferences about the posterior distribution is to obtain samples of $p(\theta|y)$.

Suppose a sample $\theta^1, \ldots, \theta^k$ of $p(\theta|y)$ is available, then, we can use the law of large numbers (if it holds) to get the approximation

$$E(h(\theta)|y) \approx \frac{1}{k}\sum_{i=1}^{k} h(\theta^i)$$

Notice that this approximation is valid regardless of the dimensionality of θ (at least in theory). Furthermore, if the quantity of interest is a transformation of θ, say $\eta = g(\theta)$ we can obtain samples of η as $g(\theta^1), \ldots, g(\theta^k)$.

Direct Sampling

In some cases it is possible to factor the posterior distribution of a multivariate parameter as the product of distributions of blocks of parameters that are simple to sample from. Thus , if θ can be split in k blocks, $(\theta_1, \ldots, \theta_k)$, then

$$p(\theta) = p(\theta_1)p(\theta_2|\theta_1), \ldots, p(\theta_k|\theta_1, \ldots, \theta_{k-1})$$

Start by sampling θ_1 and then sample recursively θ_i given the previous samples.

Importance Sampling

To approximate the expectation

$$E(g(\theta)|x) = \int_\Theta g(\theta)p(\theta|x)d\theta$$

we can use a density $h(\theta)$ that is easy to sample from. Then

$$E(g(\theta)|x) = \int_\Theta \frac{g(\theta)}{h(\theta)}p(\theta|x)h(\theta)d\theta \approx \frac{1}{m}\sum_{i=1}^{M} g(\theta^i)\omega(\theta^i), \quad \theta^i \sim H$$

where $\omega(\theta) = p(\theta|x)/h(\theta)$. Since $p(\theta|x)$ is usually known only up to a constant then

$$E(g(\theta)|x) \approx \frac{\sum_{i=1}^{M} g(\theta^i)\omega(\theta^i)}{\sum_{i=1}^{M} \omega(\theta^i)}, \quad \theta^i \sim H$$

Rejection Sampling

A rejection scheme to sample from the posterior distribution can be set in the following way.

1. Obtain a sample from a density h. This needs to be as good an approximation to $p(\theta|x)$ as possible. The ratio $p(\theta|x)/h(\theta)$ needs to be bounded above. Let M be the upper bound.

2. Accept the sample as a draw from $p(\theta|x)$ with probability

$$\frac{p(\theta|x)}{Mh(\theta)}$$

As with importance sampling, the efficiency of this scheme depends on the ability to find a good proposal distribution h.

Markov Chain Monte Carlo

The idea of iterative sampling methods based on Markov Chains (MCMC) is to build a Markov Chain which is easy to simulate and that has an equilibrium distribution equal to the distribution of interest.

In the Bayesian setting the distribution of interest is usually the posterior distribution. The method is based on choosing appropriate transition kernels $q(x, y)$ for a chain that is homogeneous, irreducible and aperiodic. This kernel has to be easy to sample from and produce a limiting distribution equal to $p(\theta|x)$.

Gibbs Sampling

The Gibbs sampler consists of iteratively sampling from the full conditionals of the parameters. Suppose that θ is made of p blocks $\theta_1, \ldots, \theta_p$. We denote the j-th sample of θ_i as θ_i^j and θ_{-i} the set of all blocks excluding θ_i.

A Gibbs Sampler proceeds as follows:

- Choose an initial configuration θ^0.
- Sample θ_1^1 from $p(\theta_1|\theta_2^0, \ldots, \theta_p^0, x)$
- Sample θ_2^1 from $p(\theta_2|\theta_1^1, \theta_3^0, \ldots, \theta_p^0, x)$
- Sample θ_i^1 from $p(\theta_i|\theta_{-i}, x)$, for $i = 3, \ldots, p$
- Once all the components of θ have been updated, cycle.

Metropolis-Hastings Sampler

The MCMC initially proposed by Metropolis and later extended by Hastings is based on choosing a transition kernel or jumping distribution $q(x, y)$ giving the probability of moving from x to y.

To obtain a chain that has $p(\theta|x)$ as the limiting distribution we cycle through the following algorithm:

- Choose and initial configuration θ^0.
- At iteration i, sample a proposal θ^* from $q(\theta^i, \theta^*)$.
- Calculate the probability

$$\alpha(\theta^i, \theta^*) = \min\left\{1, \frac{p(\theta^*|x)}{p(\theta^i|x)}\frac{q(\theta^i, \theta^*)}{q(\theta^*, \theta^i)}\right\}$$

- Let $\theta^{i+1} = \theta^*$ with probability α. Otherwise, let $\theta^{i+1} = \theta^i$.