

## Lecture Note 10: Bayesian Estimation and MCMC

### Bayes Point Estimator

Recall the basic setup: a parametric model for data:

$$z \sim P_\theta, \quad \theta \in \Theta.$$

Prior density:  $p(\theta)$ .

Posterior:  $p(\theta|z) \propto p(\theta)p(z|\theta)$ .

After calculating the posterior (either using analytic methods, or numerically using methods to be described below), we can construct various point estimators of  $\theta$ . One possibility is the posterior mean:

$$\hat{\theta} = E[\theta|z] = \int \theta p(\theta|z) d\theta.$$

More generally, we could choose a loss function  $L(\theta, \hat{\theta})$ , which measures the loss associated with choosing a point estimate  $\hat{\theta}$  when the true value is  $\theta$ . Then we would choose  $\hat{\theta}$  to minimize posterior expected loss:

$$\hat{\theta} = \arg \min_t E[L(\theta, t)|z] = \arg \min_t \int L(\theta, t) p(\theta|z) d\theta.$$

If we use squared error loss  $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ , then the solution is the posterior mean. If we use absolute error loss  $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ , then the solution can be shown to be the posterior median (that is, the median of the posterior distribution for  $\theta$ ).

Alternatively, we could report the entire posterior density  $p(\theta|z)$  and let the reader choose which functional of this distribution is most relevant.

### Interpretation of Bayes Estimators

There are different ways to interpret the Bayes estimator described above.

One perspective is to regard distributions over the parameter space as representing “beliefs” about the relative likelihood of different possible states of nature. This can be motivated axiomatically. Savage (1954)<sup>1</sup> showed that if you are a decision-maker and satisfy certain axioms of internal consistency, then you must act as if you have a utility function defined over terminal outcomes and a “subjective” probability distribution over  $\Theta$ , and maximize posterior expected utility. This extends the von-Neumann expected utility theory to settings where the probabilities (of different possible values of  $\theta$ ) are not specified as part of the decision problem.

In subjective expected utility theory, different individuals may have different subjective assessments

---

<sup>1</sup>Savage, L., 1954, *The Foundations of Statistics*, New York: Wiley, reissued in 1972 by Dover.

of  $\theta$  (hence different prior distributions). So an analysis based on a particular choice for the prior may be controversial if the reader does not like the prior used. One response would be to try to choose a relatively “uninformative” prior. Or we could report the results from a range of prior distributions.

Another perspective is to view the Bayesian point estimator as just some procedure for transforming data into estimates, and study the properties of this estimator in the same way that we study the properties of ML or GMM estimators. So we could ask: is  $\hat{\theta}$  unbiased, consistent, asymptotically normal, efficient, etc?

It turns out that in regular models with “smooth priors”, the Bayes estimator is asymptotically equivalent to the MLE. So from a large-sample “classical” perspective, the Bayes estimator has the same desirable properties as MLE. However, a dogmatic prior could have a strong influence on the estimate in practice.

### Normal Linear Regression Models

Suppose that  $Z = (Z_1, \dots, Z_n)$ , where  $Z_i = (Y_i, X_i)$ , and suppose the conditional likelihood function has the form

$$p(y|x, \theta) = \prod_{i=1}^n l(y_i|x_i, \theta),$$

where

$$l(y_i|x_i, \theta) = \mathcal{N}(x_i'\beta, \sigma^2).$$

Let

$$X = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

be the  $n \times k$  matrix of regressors and the  $n \times 1$  vector of responses, respectively.

We are going to do the analysis conditional on  $x$ . So when we talk about prior distributions, we will assume that they are independent of parameters:  $p(\beta, \sigma|x) = p(\beta, \sigma)$ .

The posterior distribution will be related to the standard LS estimator

$$b = (X'X)^{-1}X'y$$

and the typical estimator of  $\sigma^2$  given by

$$s^2 = \frac{1}{n-k} \sum_{i=1}^n (y_i - x'_i b)^2.$$

$\chi^2$  Distribution: If

$$w_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \quad \text{for } j = 1, \dots, \nu,$$

then

$$w \equiv \sum_{j=1}^{\nu} w_j^2 \sim \chi_{\nu}^2.$$

The  $\chi^2$  distribution has a density function

$$f_{\chi_{\nu}^2}(w) = c \cdot w^{\frac{\nu-2}{2}} \exp\left(-\frac{1}{2}w\right) I_{(0,\infty)}(w) \quad (1)$$

Here  $c$  is a constant, which does not depend on  $w$ , such that (1) integrates to 1.  $I_{(0,\infty)}(w)$  is the indicator function, equal to 1 if  $w \in (0, \infty)$  and equal to 0 otherwise.

We will start by examining a special prior distribution, which leads to very familiar results, and then extend the analysis to a slightly more general prior distribution. Let  $\tau = \sigma^{-2}$ , and suppose that the prior for  $\beta, \tau$  can be written as

$$p(\beta, \tau) = p_1(\tau)p_2(\beta|\tau),$$

where

$$p(\beta|\tau) \propto 1$$

and

$$p(\tau) \propto \frac{1}{\tau}.$$

Then the posterior distribution has the following form:

$$p(\beta, \tau|z) = p(\beta|\tau, z)p(\tau|z),$$

where the conditional posterior distribution of  $\beta$  is multivariate normal:

$$\beta|\tau, z \sim \mathcal{N}(b, \sigma^2(X'X)^{-1});$$

and the marginal posterior distribution of the precision  $\tau$  is scaled chi-square:

$$\tau|z \sim \frac{\chi_{(n-k)}^2}{(n-k)s^2}.$$

That is, the posterior distribution of  $\tau$  is the distribution of a chi-square  $(n-k)$  random variable divided by  $(n-k)s^2$ . Since the chi-square distribution with  $\nu$  degrees of freedom has mean  $\nu$ , the posterior mean of  $\tau$  is  $s^{-2}$ .

By the properties of the multivariate normal distribution, any linear combination of elements of  $\beta$

has a conditional normal distribution: for any  $a \in \mathbb{R}^k$ ,

$$a'\beta|\tau, z \sim \mathcal{N}(a'b, \sigma^2 a'(X'X)^{-1}a).$$

We can marginalize out the  $\tau$ :

$$a'\beta|z \sim a'b + s (a'(X'X)^{-1}a)^{1/2} t(n-k),$$

where  $t(n-k)$  stands for a  $t$  distribution with  $(n-k)$  degrees of freedom. This means that the marginal distribution of  $a'\beta$  given the data  $z$  has the same distribution as a  $t(n-k)$  random variable, scaled by  $s (a'(X'X)^{-1}a)^{1/2}$ , and recentered at  $a'b$ .

Notice that the previous two displays will be very similar for  $(n-k)$  large, suggesting that if the sample size is large (relative to the number of regressors), replacing  $\sigma^2$  with its estimate  $s^2$  and ignoring the uncertainty about the variance estimate will not greatly change the posterior distribution for  $\beta$ .

#### Prior and Posterior Distributions: Version 2

Gamma Distribution: The gamma distribution  $\mathcal{G}(x|\alpha, \beta)$  with shape parameter  $\alpha > 0$  and inverse scale parameter  $\beta > 0$  has density

$$p_g(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) I_{(0,\infty)}(x).$$

Notice that the  $\chi_\nu^2$  distribution is a special case of the gamma distribution with  $\alpha = \nu/2$ ,  $\beta = 1/2$ .

Suppose that the prior distribution for  $\beta, \tau$  has

$$\tau \sim \mathcal{G}(\tau|a_1, a_2),$$

and

$$\beta|\tau \sim \mathcal{N}(\beta_0, \tau^{-1}\Omega),$$

where  $\beta_0$  is a given  $k \times 1$  vector and  $\Omega$  is a given  $k \times k$  positive-definite symmetric matrix. Notice that letting  $a_1 \rightarrow 0$  and  $a_2 \rightarrow 0$  gives our earlier prior density for  $\tau$ , and letting  $\Omega \rightarrow 0$  gives our earlier prior density for  $\beta$ . Then the posterior distribution is given by

$$\tau|z \sim \mathcal{G}(a_1 + \frac{1}{2}n, a_n)$$

$$\beta|\tau, z \sim \mathcal{N}(\tilde{\beta}, \sigma^2(X'X + \Omega^{-1})^{-1}),$$

where

$$\tilde{\beta} = (\Omega^{-1} + X'X)^{-1}(\Omega^{-1}\beta_0 + X'y),$$

$$a_n = a_2 + \frac{1}{2}(n - k)s^2 + \frac{1}{2}(b - \beta_0)' \Omega^{-1} (X'X + \Omega^{-1})^{-1} X'X (b - \beta_0).$$

## Probit Model

Suppose we have a binary outcome and specify a probit model

$$\Pr(Y_i = 1|x) = \Phi(x_i'\beta),$$

where  $\Phi$  is the cumulative distribution function of the standard normal distribution:

$$\Phi(b) = \int_{-\infty}^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt,$$

One appealing feature of the probit setup is that it generalizes well to multinomial and multivariate models. The (conditional) likelihood function for the probit model can be written as:

$$p(y|x, \beta) = \prod_{i=1}^n \Phi(x_i'\beta)^{y_i} (1 - \Phi(x_i'\beta))^{1-y_i},$$

where  $y = (y_1, \dots, y_n)'$  and  $x = (x_1, \dots, x_n)'$ . This is not too hard to calculate on a computer (for example, in Matlab there is a function called `erf` which evaluates  $\Phi$ ), although for generalizations of the probit model, simply calculating the likelihood can be very difficult. Directly integrating the likelihood with respect to  $\beta$  to obtain the posterior distribution is still little difficult, because of the normalizing constant factor.

We will develop a *simulation-based method* for generating draws from the posterior distribution of  $\beta$ .

### Data Augmentation for the Probit Model

The probit model can be rewritten in a *Latent Variable Form*:

$$\begin{aligned} \epsilon_i | X &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1). \\ y_i^* &= x_i'\beta + \epsilon_i, \\ y_i &= \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases} \end{aligned}$$

This is an equivalent model, in the sense that for any given  $\beta$  and any  $X$  it yields the same

distribution for  $y$ . To see this, note that

$$\begin{aligned}
Pr(y_i = 1|x, \beta) &= Pr(y_i^* > 0|x, \beta) \\
&= Pr(x'_i\beta + \epsilon_i > 0|x, \beta) \\
&= Pr(\epsilon_i > -x'_i\beta|x, \beta) \\
&= Pr(\epsilon_i < x'_i\beta|x, \beta) \\
&= \Phi(x'_i\beta).
\end{aligned}$$

It will be useful to work with the joint distribution of  $\beta$  and the latent variables  $y^* \equiv (y_1^*, \dots, y_n^*)$ . We can write

$$\begin{aligned}
p(\beta, y^*|y, x) &\propto p(\beta, y^*|x)p(y|y^*, x, \beta) \\
&\propto p(\beta|x)p(y^*|\beta, x)p(y|y^*, x, \beta)
\end{aligned}$$

The other two terms on the right hand side of the preceding display have a product form, so we can write

$$p(\beta, y^*|y, x) \propto p(\beta) \prod_{i=1}^n p(y_i^*|x_i, \beta) \prod_{i=1}^n p(y_i|y_i^*, x_i, \beta)$$

Note that

$$p(y_i^*|x_i, \beta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i^* - x'_i\beta)^2\right)$$

and

$$p(y_i|y_i^*, x_i, \beta) = 1(y_i = 1)1(y_i^* > 0) + 1(y_i = 0)1(y_i^* \leq 0).$$

Putting all this together, we get:

$$p(\beta, y^*|y, x) \propto p(\beta) \prod_{i=1}^n \{1(y_i = 1)1(y_i^* > 0) + 1(y_i = 0)1(y_i^* \leq 0)\} \phi(y_i^*|x'_i\beta, 1),$$

where  $\phi(\cdot|\mu, \sigma^2)$  denotes the normal density function with mean  $\mu$  and variance  $\sigma^2$ .

Suppose that we could actually observe the latent  $y_i^*$ . Let  $y^* = (y_1^*, \dots, y_n^*)'$ . Then the posterior distribution for  $\beta$  would be easy to calculate. If we use a constant prior for  $\beta$ , we could use the results for the normal linear model to obtain

$$\beta|y^*, X \sim \mathcal{N}(b^*, (X'X)^{-1}),$$

where

$$b^* = (X'X)^{-1}X'y^*.$$

Formally, we can write

$$\begin{aligned} p(\beta|y^*, y, x) &= \frac{p(\beta, y^*|y, x)}{p(y^*|y, x)} \\ &\propto p(\beta) \prod_{i=1}^n \phi(y_i^*|x_i'\beta, 1) \end{aligned}$$

which gives the desired result.

Conversely, suppose we knew  $\beta$ . Then it would be easy to draw for the latent variables  $y^*$  from their distribution conditional on  $\beta$  and  $x$ , which is normal with mean  $x_i'\beta$  and variance 1. However, if we also condition on  $y_i$ , we need to take into account this additional source of information. Heuristically, observing  $y_i = 1$  tells us that  $y_i^* > 0$  and observing  $y_i = 0$  tells us that  $y_i^* \leq 0$ . Formally, write

$$p(y_i^*|\beta, x_i, y_i) \propto \begin{cases} \phi(y_i^*|x_i'\beta, 1)1(y_i^* > 0) & \text{if } y_i = 1 \\ \phi(y_i^*|x_i'\beta, 1)1(y_i^* \leq 0) & \text{if } y_i = 0 \end{cases}$$

Thus, if  $y_i = 1$ , the distribution of  $y_i^*$  is a normal distribution with mean  $x_i'\beta$  and variance 1, truncated so that  $y_i^*$  is strictly positive. Likewise, if  $y_i = 0$  then the distribution is truncated so that  $y_i^*$  is negative.

The idea behind data augmentation is to augment the observed data with the latent data, and iterate between these two distributions, generating a sequence of draws for  $\beta$  and  $y^*$ . Start by setting  $\beta^{(1)} = (0, \dots, 0)'$  or some other vector of numbers. (We will use superscript to denote the iteration number.) Then draw:

$$\begin{aligned} y^{*(1)} &\sim p(y^*|\beta^{(1)}, x, y) \\ \beta^{(2)} &\sim p(\beta|y^{*(1)}, x, y) \\ y^{*(2)} &\sim p(y^*|\beta^{(2)}, x, y) \\ &\vdots \end{aligned}$$

At each draw for  $\beta$ , we substitute the most recent draw for the latent  $y^*$ , and likewise in the draws for  $y^*$ .

This is a special type of Gibbs sampler, which involves drawing for parameters and latent variables. It can then be shown that

$$(\beta^{(J)}, y^{*(J)}) \xrightarrow{d.} p(\beta, y^*|y, x), \quad \text{as } J \rightarrow \infty. \quad (2)$$

In addition, for any integrable function  $g(\beta, y^*)$ ,

$$\frac{1}{J} \sum_{j=1}^J g(\beta^{(j)}, y^{*(j)}) \xrightarrow{\text{a.s.}} \int g(\beta, y^*) p(\beta, y^*|y, x) d\beta dy^*, \quad \text{as } J \rightarrow \infty. \quad (3)$$

The average on the left of expression (3) is a “time” or *ergodic* average. Notice that the result is similar to a law of large numbers, but that we are not generating the draws for  $\beta, y^*$  independently. In fact, we are not even generating them from the “correct” distribution, although according to expression (2) the distribution approaches the correct distribution eventually.

Since according to (2) the posterior distribution of  $\beta, y^*$  converges jointly, it is also true that the marginal distributions converge, so

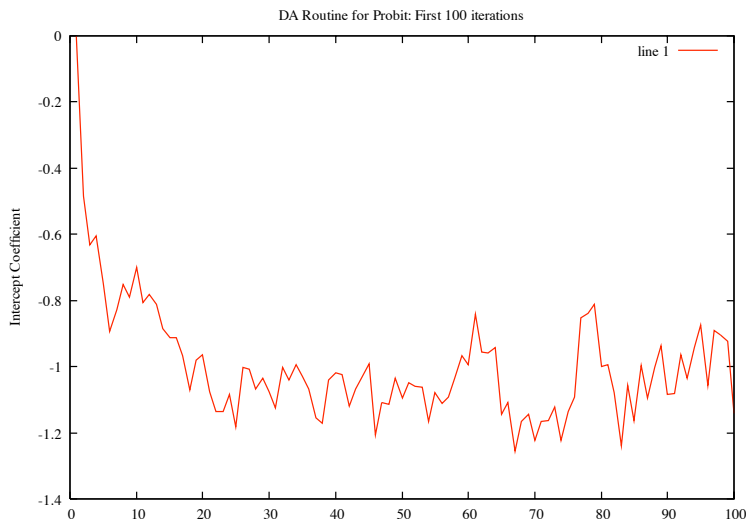
$$\beta^{(j)} \xrightarrow{d.} p(\beta|y, x), \quad \text{as } j \rightarrow \infty.$$

Also, it follows from (3) that for an integrable function  $h(\beta)$ ,

$$\frac{1}{J} \sum_{j=1}^J h(\beta^{(j)}) \xrightarrow{\text{a.s.}} \int h(\beta)p(\beta|y, x)d\beta, \quad \text{as } J \rightarrow \infty.$$

So we can run a single chain and get, for example, the posterior mean of  $\beta$ .

Here is a sample of what the output from such a chain looks like:



As you can see, the chain appears to converge fairly quickly to a distribution centered around -1.1.

In practice, it is a good idea to run multiple chains in parallel, with different starting values. Then we can plot the sequence of draws for  $\beta$  for each chain, and get a visual sense for whether the chains have converged.