Note Set 1 – Bayesian Methods

1.1 - Introduction

1.1.1 – Decision Theoretic Approach to Estimation

There are two ways to view Bayesian estimation. The first approach starts with of idea of treating statistical estimation as a decision theoretic process. An individual must choose an estimator $\hat{\theta}$. He makes this choice by maximizing his expected utility (or by minimizing his expected loss). We denote the loss function by, $L(\theta,\theta')$, where this denotes the loss from choosing the estimator θ when the truth is θ'. The individual has prior belief $\pi(\theta)$ which he updates using a data vector x to form a posterior $f(\theta \mid x)$. We then selects $\hat{\theta}$ by minimizing his expected loss, i.e.,

$$\hat{\theta} = \arg\max_{\theta} \int_{\theta'} L(\theta,\theta') f(\theta' \mid x) d\theta'$$

This estimator has certain nice properties even in finite samples--for an individual with the same loss function and prior on the parameters, the parameter estimate is optimal in a decision theoretic sense.

If we report the entire posterior distribution $f(\theta'|x)$, we would provide any individual with the same prior (but not necessarily the same loss function) with enough information to determine his optimal decision. This approach too has a nice finite sample property. Any individual with the same prior can use posterior to come to an optimal point estimate from a decision theoretic sense.

1.1.2 – Estimation Using Integration

The other way to view Bayesian estimation is an alternative approach to generate estimators with nice properties. In particular, Bayesian estimators are an alternative to maximum likelihood estimators, M-estimators, Generalized Method of Moments estimators, etc., and have nice properties including consistency, asymptotic normality, and efficiency. These are our course, the same properties obtained by maximum likelihood estimators in parametric models. The advantages of the Bayesian estimation is that in some cases, the Bayesian estimator will be easier to implement or more computationally efficient than alternative estimators (or course, there are cases where the Bayesian estimator will be harder to implement or less computationally efficient).

Unlike most other estimation methods, Bayesian estimation does not rely on optimization, and instead relies on integration. Bayesian estimation is likely to be a preferred alternative when the objective function obtain from optimization-based

estimator are discontinuous, or when numerical integration is required to compute the objective function. Bayesian estimators are likely to be less effective in models that are non-linear in parameters, in which case efficient methods of sampling from the posterior distribution may not be available.

1.1.3 – Assumptions

It is important to realize that while Bayesian estimator achieve some nice properties that other estimators do not (minimization of loss in finite samples), they require additional assumptions (the loss function and prior distribution are correct). In academic work, we often strive to convey our results to others in which case these constraints are unattractive (i.e. our audience may not share our loss function or prior distribution).

It would be desirable to develop conditions under which our estimators would have desirable decision theoretic properties that would apply to all members of our audience. These conditions require employing large sample approximations. Specifically, Bayesian estimators are consistent, asymptotically normal, efficient, and converge asymptotically to the maximum likelihood estimator.

These results have a number of important implications. First, in large samples, all Bayesian estimators and the maximum likelihood estimator converge to the same point (regardless of the prior and loss function chosen). We also have that $\sqrt{N}(\hat{\theta}-\theta_0)$ will have the same large sample distribution (regardless of the prior and loss function chosen). These results imply that in large samples, Bayesian point estimation will lead to the

correct decision, regardless of the choice of loss function and prior distribution. The same will hold for any consistent estimator. The results also imply that in large samples, Bayesian inferences will be correct (regardless of the prior and loss function chosen). The same will hold for any consistent, asymptotically normal, and efficient estimator, meaning that classical estimators can be interpreted as large sample approximations to the posterior.

Table 1.1 – Comparison of Maximum Likelihood and Bayesian Estimators

	MLE	**Bayesian Posterior**	**Bayesian Point Estimator**
Nice Finite Sample Properties	In general, Nothing	Posterior can be used to make correct decisions if prior is correct	Correct decision if prior and loss function are correct
Nice Large Sample Properties	1. Consistent 2. Asymptotically Normal 3. Efficient 4. Correct Decisions	1. Consistent 2. Asymptotically Normal 3. Efficient 4. Correct Decisions (regardless of whether prior is correct)	1. Consistent 2. Asymptotically Normal 3. Efficient 4. Correct Decisions (regardless of whether prior and loss function are correct)

We summarize these results in Table 1. In general maximum likelihood estimators do not have any nice properties in finite samples. Bayesian inferences are correct in finite sample provided that the prior is "correct". Bayesian point estimators are optimal in finite sample provided that the prior and loss function are both "correct". In large samples, all three approaches lead to consistency, asymptotic normality, efficiency, and correct inferences (from a decision theoretic perspective). The differences between these

estimators in large samples are their ease of implementation and computational efficiency.

In finite samples, Bayesian estimators have some nice properties which come at a cost. You can think of Bayesian estimators as lying along a continuum of estimators that differ in the number of nice properties they achieve and the number of assumptions necessary to ensure that these nice properties hold. At one end of the spectrum are nonparametric estimators, which achieve few nice properties, but require very few assumptions to achieve those properties. Towards the middle are maximum likelihood estimators which require a parametric model, but achieve consistency and efficiency in large samples. At the other end are Bayesian estimators which achieve some nice finite sample properties, but require additional assumptions. Most of the time, the best estimator to use should fall somewhere in the middle of this continuum, using either parametric or semi-parametric methods. Only in the most data rich tasks, or tasks that involve low dimensional analysis, should you consider using fully nonparametric methods. And only in the situations of the most limited data should you consider using Bayesian methods to make finite sample inferences. Using Bayesian estimators for large sample inference is more generally appropriate, provided that the Bayesian estimator provides computational advantages over the MLE.

1.1.4 – Posterior Distributions and Loss Functions

Let $\pi(\theta)$ denote the prior distribution of θ and let $L(\theta \mid x) = \prod_{n=1}^{N} f(x_n; \theta)$ denote the likelihood function. The posterior distribution of θ is derived using Bayes rule,

$$f(\theta \mid x) = \frac{f(x,\theta)}{f(x)} = \frac{L(\theta \mid x)\pi(\theta)}{f(x)}$$

Recall that a Bayesian point estimator was defined by,

$$\hat{\theta} = \arg\min_{\theta} \int_{\theta'} L(\theta,\theta') f(\theta' \mid x) d\theta'$$

Here, the loss function should be minimized when $\theta = \theta'$ (i.e. loss is minimized when the estimate equals the truth. Two important choices include the quadratic loss function,

$$L(\theta,\theta') = -\tfrac{1}{2}(\theta - \theta')W(\theta - \theta')$$

where W is a symmetric positive definite matrix, and,

$$L(\theta,\theta') = \sum_{k=1}^{K} w_k \mid \theta_k - \theta_k' \mid$$

where $w_k > 0$. In the first case, we can use first order conditions to obtain,

$$\hat{\theta} = \int_{\theta'} \theta' f(\theta' \mid x) d\theta'$$

In other words, when the loss function is quadratic, the Bayesian point estimator is the posterior mean. Similarly, we have that the absolute value loss function leads to the posterior dimension-by-dimension median as the Bayesian point estimator.

1.1.5 – Estimating a Proportion

Suppose that $x_n \sim Bernouli(\theta)$. We have a likelihood of,

$$L(x \mid \theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} = \theta^{y}(1-\theta)^{N-y}$$

where $y = \sum_{n=1}^{N} x_n$. If we assume a prior distribution of $\theta \sim U[0,1]$, we have,

$$f(\theta \mid x) = \frac{\theta^y (1-\theta)^{N-y}}{f(x)}$$

Recall that the Beta distribution is given by,

$$f_X(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x^{\beta-1})$$

We have,

$$f(x) = \int_{\theta=0}^{1} \theta^y (1-\theta)^{N-y} d\theta = \frac{B(y+1,N+1-y)}{B(y+1,N+1-y)} \int_{\theta=0}^{1} \theta^{(y+1)-1}(1-\theta)^{(N-y+1)-1} d\theta = B(y+1, N-y+1)$$

$$f(\theta \mid x) = \frac{\theta^y (1-\theta)^{N-y}}{B(y+1, N-y+1)} \sim Beta(y+1, N-y+1)$$

where,

$$E[\theta \mid x] = \frac{y+1}{N+2}$$

The result indicates that with the uniform prior distribution, the estimate is biased towards $\frac{1}{2}$, but that this bias diminishes as the sample size increases.

We can consider a more flexible prior distribution. Let use suppose that we have a Beta prior distribution of,

$$\pi(\theta;\alpha_{prior},\beta_{prior}) = \frac{1}{B(\alpha_{prior},\beta_{prior})} \theta^{\alpha_{prior}-1}(1-\theta^{\beta_{prior}-1})$$

We also have,

$$f(x) = \int_{\theta=0}^{1} \theta^y (1-\theta)^{N-y} \frac{1}{B(\alpha_{prior},\beta_{prior})} \theta^{\alpha_{prior}-1}(1-\theta^{\beta_{prior}-1}) d\theta$$

$$= \frac{1}{B(\alpha_{prior},\beta_{prior})} \int_{\theta=0}^{1} \theta^{\alpha_{prior}+y-1}(1-\theta)^{\beta_{prior}+N-y-1} d\theta$$

$$= \frac{1}{B(\alpha_{prior},\beta_{prior})} \frac{B(\alpha_{prior}+y,\beta_{prior}+N-y)}{B(\alpha_{prior}+y,\beta_{prior}+N-y)} \int_{\theta=0}^{1} \theta^{\alpha_{prior}+y-1}(1-\theta)^{\beta_{prior}+N-y-1} d\theta$$

$$= \frac{B(\alpha_{prior}+y,\beta_{prior}+N-y)}{B(\alpha_{prior},\beta_{prior})}$$

We obtain,

$$f(\theta \mid y) = \frac{\theta^{y}(1-\theta)^{N-y}\,\frac{1}{B(\alpha_{prior},\beta_{prior})}\,\theta^{\alpha_{prior}-1}(1-\theta^{\beta_{prior}-1})}{\frac{B(\alpha_{prior}+y,\beta_{prior}+N-y)}{B(\alpha_{prior},\beta_{prior})}}$$

$$= \frac{\theta^{y+\alpha_{prior}-1}(1-\theta^{N-y+\beta_{prior}-1})}{B(\alpha_{prior}+y,\beta_{prior}+N-y)}$$

As a result, we have,

$$\theta \mid Y \sim Beta(y+\alpha_{prior}, N-y+\beta_{prior})$$

Notice that the posterior is Beta just like the prior. In this case, we say that the Beta distribution is a conjugate prior. In addition to the uniform distribution, conjugate priors often lead to the simplest Bayesian estimators. Now, recall that for a Beta random variable, $E[X] = \frac{\alpha}{\alpha+\beta}$. We therefore have,

$$E[\theta \mid x] = \frac{y+\alpha_{prior}}{N+\alpha_{prior}+\beta_{prior}}$$

What we have then is that the posterior mean is adjusted by a factor dependent on the prior. This factor diminishes in importance as the sample size becomes large.

1.1.6 – Estimating the Mean of a Normal Distribution (Known Variance)

Suppose that $x_1, x_2, ..., x_N$ are a sample from $N(\mu, \sigma^2)$ where σ^2 is known a-priori. Suppose that we wish to find a Bayesian estimator $\hat{\mu}$ or μ, and suppose that our prior distribution is $N(\mu_{prior}, \sigma^2_{prior})$. We have that,

$$L(x \mid \mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{N} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2}$$

$$\pi(\mu;\mu_{prior},\sigma_{prior}) = \frac{1}{\tau_{prior}\sqrt{2\pi}} e^{-\frac{1}{2}(\mu-\mu_{prior})^2/\sigma^2_{prior}}$$

We can determine that the posterior distribution is given by,

$$f(\mu \mid x) = \frac{Top}{Bottom} \frac{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2} \frac{1}{\sigma_{prior}\sqrt{2\pi}} e^{-\frac{1}{2}(\mu-\mu_{prior})^2/\sigma^2_{prior}}}{\int_{\mu=-\infty}^{\infty}\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2} \frac{1}{\sigma_{prior}\sqrt{2\pi}} e^{-\frac{1}{2}(\mu-\mu_{prior})^2/\sigma^2_{prior}}\right] d\mu}$$

Considering just the top part, we have,

$$Top = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \frac{1}{\sigma_{prior}\sqrt{2\pi}} e^{-\frac{1}{2}\left\{\mu^2\left[\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right]-2\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]\mu+\frac{\overline{x^2}}{\sigma^2/N}+\frac{\mu^2_{prior}}{\sigma^2_{prior}}\right\}}$$

Working with such expressions quickly becomes intractable, so we define,

$$c_1 = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \frac{1}{\sigma_{prior}\sqrt{2\pi}} e^{-\frac{1}{2}\left\{\frac{\overline{x^2}}{\sigma^2/N}+\frac{\mu^2_{prior}}{\sigma^2_{prior}}\right\}}$$

We have,

$$Top = c_1 e^{-\frac{1}{2}\left\{\mu^2\left[\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right]-2\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]\mu\right\}} = c_1 e^{-\frac{1}{2\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}}\left\{\mu^2-2\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]\mu\right\}}$$

The next step involves completing the square. We have,

$$Top = c_1 e^{-\frac{1}{2\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}}\left\{\left(\mu-\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]\right)^2-\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-2}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]^2\right\}}$$

Defining,

$$c_2 = e^{\frac{1}{\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}}\left\{\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-2}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]^2\right\}}$$

we have,

$$Top = c_1 c_2 e^{-\frac{1}{\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}}\left\{\left(\mu-\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]\right)^2\right\}}$$

and,

$$Bottom = c_1 c_2 \int_{\mu=-\infty}^{\infty} \left[e^{-\frac{1}{\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}}\left\{\left(\mu-\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]\right)^2\right\}} \right] d\mu$$

we further have that,

$$\int_{\mu=-\infty}^{\infty} \left[e^{-\frac{1}{\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}}\left\{\left(\mu-\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]\right)^2\right\}} \right] d\mu = \frac{1}{\sqrt{2\pi\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}}}$$

which implies that,

$$f(\mu \mid x) = \frac{e^{-\frac{1}{\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}}\left\{\left(\mu-\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right]\right)^2\right\}}}{\sqrt{2\pi\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}}}$$

Hence, we have,

$$\mu \mid x \sim N(\left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right], \left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1})$$

A Bayesian point estimator involves finding a value $\hat{\mu}$ that minimizes the expected loss. If we choose a quadratic loss function, then the mean of the posterior will minimize the expected loss. Hence, we have,

$$\hat{\mu} = \left(\frac{1}{\sigma^2/N}+\frac{1}{\sigma^2_{prior}}\right)^{-1}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\sigma^2_{prior}}\right] = \frac{\bar{x}+\frac{\sigma^2/N}{\sigma^2_{prior}}\mu_{prior}}{1+\frac{\sigma^2/N}{\sigma^2_{prior}}}$$

Notice that as $N \to \infty$, we have that $\frac{\sigma^2/N}{\sigma^2_{prior}} \to 0$. Hence, $\hat{\mu} \xrightarrow{prob.} \bar{x} \xrightarrow{prob.} \mu_0$. Next, as $N \to \infty$, we find $\left(\frac{1}{\sigma^2/N} + \frac{1}{\sigma^2_{prior}}\right)^{-1} \to \sigma^2 / N$. Hence, the Bayesian posterior mean has the same asymptotic distribution as the sample mean.

In this case, we were able to derive a closed form solution for the Bayesian estimator. This was quite tedious, but is made easier by the fact that we can ignore constant terms. From now on, we will ignore the denominator in deriving posterior distributions.

1.2 – Numerical Integration

In order to compute quantities of interest in Bayesian estimator, we must be able to compute integrals of the following form,

$$\int_\theta h(\theta) f(\theta \mid x) d\theta = \frac{\int_\theta h(\theta) L(\theta \mid x) \pi(\theta) d\theta}{\int_\theta L(\theta \mid x) \pi(\theta) d\theta}$$

Traditional numerical integration is not particularly useful because in most interesting problems, the dimensionality of θ will be large. Instead, we will rely on Monte Carlo integration techniques. These techniques rely on drawing a random sample $\{\theta_{(r)}\}_{r=1}^R$ and approximating,

$$\int_\theta h(\theta) f(\theta \mid x) d\theta \approx \frac{1}{R} \sum_{r=1}^R h(\theta_{(r)})$$

In this section, we develop techniques for drawing a random sample from $f(\theta \mid x)$.

1.2.1 – Using the Inverse CDF

Through this section, we assume that we have the technology to draw $Uniform(0,1)$ random deviates, and that we have access to utilities to draw deviates from the most common distributions. For this reason, we don't focus on drawing from the normal distribution, gamma distribution, etc.

Suppose that we would like to draw a random variable X from the cdf F_X and suppose that U is a uniform random deviate. Then we can show that $X = F_X^{-1}(U) \sim F_X$. To see that this is the case,

$$\Pr(X \le x) = \Pr(F_X^{-1}(U) \le x) = \Pr(U \le F_X(x)) = F_X(x)$$

For example, the truncated normal distribution has the pdf,

$$f_X(x;\mu,\sigma^2,a,b) = \frac{\phi(x;\mu,\sigma^2)}{\Phi(b;\mu,\sigma^2) - \Phi(a;\mu,\sigma^2)}$$

We can determine that the cdf is given by,

$$F_X(x;\mu,\sigma^2,a,b) = \frac{\Phi\left(\frac{x-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

We can invert these to obtain,

$$F^{-1}(x) = \sigma\Phi^{-1}\left[x[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)] + \Phi\left(\frac{a-\mu}{\sigma}\right)\right] + \mu$$

Hence, we can draw truncated normal random deviates using,

$$X = \sigma\Phi^{-1}\left[[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)]U + \Phi\left(\frac{a-\mu}{\sigma}\right)\right] + \mu$$

where U is a uniform random deviate.

1.2.2 – Rejection Sampling

Consider again the problem of drawing from the truncated normal distribution. Consider the following algorithm:

1. General a draw from $X \sim N(\mu, \sigma^2)$
2. If $a < X < b$, then accept X. Otherwise, go back to step 1.

This method is actually much more general than this. Consider any $g(x)$ such that $f(x) \leq Mg(x)$ for all x for some specific $M > 0$. Consider the following algorithm,

1. Sample X from $g(x)$ and a uniform random deviate U
2. If $U < \frac{f(X)}{Mg(X)}$, then accept X. Otherwise, go to step 1.

To use this technique, we need to develop an appropriate $g(x)$. We could select Mg to be very big, but this would mean that we would often be rejecting and this would lead to very inefficient sampling. We could select Mg to be equal to f, but now we are just back at the original problem where we don't know how to sample from f. Instead, we want to choose a g and M such that Mg is very close to f and g is easy to sample from. One approach that is used for log-concave densities. This approach generates g using a piecewise linear approximation to $\log f$. The "adaptive" part comes from the fact that the linear approximation is chosen so that not too many samples are rejected.

1.2.3 – Importance Sampling

Suppose that we would like to draw from the density $f(x)$, but cannot easily obtain such draws. Suppose, however, that we can easily obtain draws from $g(x)$. We can then draw a random sample from $x_{(1)}, x_{(2)}, .., x_{(R)}$ from g and weight each observation by $\frac{f(x)}{g(x)}$. We can approximate the integral,

$$\int_x h(x) f(x) dx \approx \frac{1}{R} \sum_{r=1}^{R} h(x_{(r)}) \frac{f(x_{(r)})}{g(x_{(r)})}$$

Notice that the Law or Large Numbers implies that,

$$\frac{1}{R} \sum_{r=1}^{R} h(x_{(r)}) \frac{f(x_{(r)})}{g(x_{(r)})} \xrightarrow{prob.} E\left[h(x_{(r)}) \frac{f(x_{(r)})}{g(x_{(r)})} \right] = \int_x h(x) \frac{f(x)}{g(x)} g(x) dx = \int_x h(x) f(x) dx$$

justifying our approach.

1.2.4 – Markov Chain Monte Carlo

Markov Chain Monte Carlo algorithms are a class of algorithms for Monte Carlo integration. Our goal is to generate of sequence of draws $\{X_t\}_{t=1}^{T}$ form a target distribution $F_X(x)$. Unlike any of the techniques we have studied so far, Markov Chain Monte Carlo algorithms generate dependent draws. These draws are not independent draws from $F_X(x)$, but form a sequence of dependent draws whose stationary distribution is $F_X(x)$. The Law of Large Numbers for dependent random variables

(White, 1984) then indicates that the moments of the Markov Chain $\{X_t\}_{t=1}^{T}$ will converge to the moments of the stationary distribution, i.e.,

$$\frac{1}{T}\sum_{t=1}^{T} h(X_t) \xrightarrow{prob.} \int h(x)dF_X(x)$$

Markov Chain Monte Carlo methods are primarily used in Bayesian estimation. Unlike many of techniques we have studied, it has the advantage of allowing use to sample from a density without knowing the constant of integration.[1]

Metropolis Hastings is a very general algorithm for performing MCMC integration (and all of the algorithms we will study in this subsection are special cases of Metropolis Hastings). We want to generate a Markov Chain $X_1, X_2, ..., X_T$ whose stationary distribution is $f_X(x)$. Given a current draw, x, we generate a new draw x' from the distribution $g(x'|x)$. We take a draw U from the uniform distribution. We 'accept' the step x' is $U < \frac{f(x)g(x|x')}{f(x')g(x'|x)}$. Then, we have,

$$x_{t+1} = \begin{cases} y_t, & U_t < \frac{f(x_t)g(x_t|y_t)}{f(y_t)g(y_t|x_t)} \\ x_t, & otherwise \end{cases}$$

where $U_t \sim Uniform(0,1)$ and $y_t \sim g(y|x_t)$.

We already know the conditional distribution, $f(x_{t+1}|x_t)$. We would like to characterize the stationary distribution, which is characterized by,

$$f(x') = \int_x f(x',x)dx = \int_x f(x'|x)f(x)dx$$

[1] Rejection sampling can also be performed without knowing the constant of integration.

Proving that the Metropolis Hastings algorithms works requires two steps. The first requires show that the conditions for the Dependent Law of Large numbers are satisfied. We can accomplish by showing that the Markov Chain we generate is stationary and ergodic. Then, we have to show that the stationary distribution characterized above is equal for $f_X(x)$.

1.2.5 – The Random Walk Sampler

The random walk sampler is a special case of the Metropolis Hastings sampler where $g(y \mid x) = g(y - x)$. A typical choice might by $g(y - x) \sim N(0, D)$ where D is a diagonal matrix. The tricky part about this is picking the tuning parameters, D. To simplify things, we might try $D = \sigma^2 I$, but we still have one tuning parameter to select. The performance of the sampler will depend crucially on the value of the tuning parameters. There are different rules that we can use here. A rule of thumb is to select σ such that between 20% and 40% of draws are accepted.

It would be quite tempting to select the tuning parameter σ on the basis of previous iterations. This is however, not a good idea since it can disturb the stationary distribution to which the process converges. It is valid to set the tuning parameter based on a previous run. We could try to automatically select σ using a pre-burnin period. We run the chain for a small number of iterations. If the acceptance rate is too high, we increase σ. If the acceptance rate it too low, then we decrease σ. Only then do we start the burnin period, to make sure that we allow the chain to converge to the stationary

distribution of the MCMC chain, and not the stationary distribution of the alerted chain $\{X_t, \sigma_t\}$ where σ_t is updated according to some rule.

1.2.6 – The Gibbs Sampler

The Gibbs sampler is a special case of the Metropolis Hastings algorithm. Let us consider the two dimensional case for simplicity. Suppose that the conditional distribution as $f_{X|Y}(x \mid y)$ and $f_{Y|X}(y \mid x)$. In the Gibbs samples, we set $x' \sim f(x \mid y)$ and $y' \sim f(y \mid x')$. This corresponds to $g(x', y' \mid x, y) = f(x' \mid y) f(y' \mid x')$. The Gibbs sampler is special (and often preferred) because tuning parameters are not necessary.

The drawback of the Gibbs sampler is that we must somehow be able to sample from the conditional distributions. Often, when conjugate prior distributions are used, it is possible to directly sampler form the conditional distributions. When this is not possible, Gibbs sampling allows us to reduce the problem of sampling form a multidimensional distribution to the problem of sampling from a few univariate distributions.

On approach is to combine the Gibbs sampler with one of the algorithms we covered previously. The obvious choices are (i) inversion of the cdf, (ii) rejection sampling, (iii) and the random walk sampler. Consider the inversion algorithm for sampling from a cdf $F_X(x)$. We can sample $U \sim Uniform(0,1)$ and then solve $F_X(X) = U$. This requires that (i) we are able to compute F_X, and that we can solve the nonlinear system $F_X(x) = y$. This will generally not be an effective approach.

An alternative is to use rejection sampling, but this approach is not completely straightforward since it requires selecting a proposal distribution. Good performance will

require selecting a proposal density which does not deviate from $f_X(x)$ by too much. A Third approach is to sample from the univariate distributions using the random walk sampler in one dimension. This method is the easiest to implement, but we come back to the same problem of having to select tuning parameters.

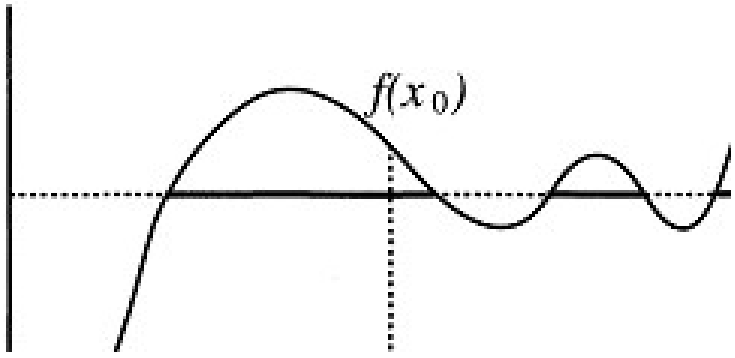
1.2.7 – Slice Sampling

Slice sampling was a technique designed to avoid the need to select tuning parameters (Neal, 2003). The idea behind slice sampling is sample uniformly from the area under $f_X(x)$, keeping only the x coordinate of the draw. This can be accomplished by sampling two uniform random variables.

If we want to sample $X \sim f_X(x)$, we introduce an auxiliary variable Y. Given some initial draw X_{t-1}, we draw $Y_t \sim Uniform(0, f_X(X_{t-1}))$. Then, we draw $X_t \sim Uniform(f_X^{-1}[0, Y_t])$. This generate a Markov Chain $\{X_t, Y_t\}$ where X_t has a stationary distribution of $f_X(x)$.

The most involved step is determining the set $f_X^{-1}[0, Y_t]$. This problem is illustrated in Figure 6.1. The problem is that the set may be potentially disconnected, so even in one dimension, we require finding all solution to $f_X(x) = Y_t$. Of course, this approach will not be effective. Neal's (2003) approach to slice sampling is find shortcuts for finding reasonable approximations of the set $f_X^{-1}[0, Y_t]$, that don't require many evaluations of f_X.

Figure 1.1 – Slice Sampling Example



1.2.8 – Hamiltonian Monte Carlo

An alternative approach (also an MCMC method) is Hamiltonian Monte Carlo. Unlike the other methods considered so far, Hamiltonian Monte Carlo generally does not use the Gibbs sampler to break up the sampling into blocks. It is however difficult to tune.

1.2.9 – Convergence Diagnostic

Consider a sequence of draws from a Markov chain $\{X_r\}_{r=1}^{R}$. If these are drawn from the stationary distribution, is should be the case that

$\bar{X}_1 = \frac{1}{R_1}\sum_{r=1}^{R_1} X_r \approx \bar{X}_2 = \frac{1}{R_2}\sum_{r=R-R_2+1}^{R} X_r$. The Geweke diagnostic checks the equality of two such means (typically $R_1 = \frac{1}{10}R$ and $R_2 = \frac{1}{2}R$. Provided that the chain is long enough, $\bar{X}_1$

and $\overline{X}_2$ are independent so that the variance of the difference is the sum of the variances. The variance of $\overline{X}_1$ and $\overline{X}_2$ must take into account the time series dependence in the series and Geweke suggests using a spectral density approach.

The Gelman and Rubin diagnostic require multiple Markov chains. Let $f_0(X)$ be the stationary distribution and assume that the start points of M chains are drawn from a distribution that is more dispersed than $f_0(X)$. Gelman and Rubin's argument is that when convergence has occurred, the across chain variance is equal to the within chain variance. The develop a statistic $\hat{R}$ which is greater than one if the across chain variance is larger and equal to one if the variances are equal. They suggest that convergence has occurred when $\hat{R} = 1.001$.

1.3 – Bayesian Estimation Using the Gibb Sampler

1.3.1 – Estimating the Mean of a Normal Distribution (Unknown Variance)

Suppose that we want to estimate σ^2 in addition to μ. Suppose that the prior distributions are given by,

$$\mu \sim N(\mu_{prior}, \tau^2_{prior})$$

$$\sigma^2 \sim Inv-\chi^2(\sigma^2_{prior}, \nu_{prior})$$

The scaled inverse chi-squared distribution is given by,

$$f(x;\nu,\sigma^2) = \frac{(\sigma^2\nu/2)^{\nu_{prior}/2}}{\Gamma(\nu/2)} \frac{e^{-\nu\sigma^2/(2x)}}{x^{1+\nu/2}}$$

We can determine that the posterior distribution is given by,

$$f(\mu,\sigma^2 \mid x) \propto e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{1}{2\tau_{prior}^2}(\mu-\mu_{prior})^2} e^{-\nu_{prior}\sigma_{prior}^2/(2\sigma^2)} \sigma^{-N-2(1+\nu_{prior}/2)}$$

Relying on some of the derivations from the previous section, we have that,

$$f(\mu,\sigma^2 \mid x) \propto e^{-\frac{1}{2}\left\{\mu^2\left[\frac{1}{\sigma^2/N}+\frac{1}{\tau_{prior}^2}\right]-2\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\tau_{prior}^2}\right]\mu+\frac{\overline{x^2}}{\sigma^2/N}+\frac{\mu_{prior}^2}{\tau_{prior}^2}\right\}} e^{-\nu_{prior}\sigma_{prior}^2/(2\sigma^2)} \sigma^{-N-2(1+\nu_{prior}/2)}$$

The posterior distribution no longer comes from a known family of distributions (and in particular, we cannot analytically compute the mean of the distribution). We note, however, that we can characterize the distribution of $\mu \mid \sigma^2, x$ and $\sigma^2 \mid \mu, x$ in closed form.

We have,

$$f(\mu \mid \sigma^2, x) \propto e^{-\frac{1}{2}\left\{\mu^2\left[\frac{1}{\sigma^2/N}+\frac{1}{\tau_{prior}^2}\right]-2\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\tau_{prior}^2}\right]\mu\right\}}$$

We can immediately guess that,

$$\mu \mid \sigma^2, x \sim N\left(\left(\frac{1}{\sigma^2/N}+\frac{1}{\tau_{prior}^2}\right)^{-1}\left[\frac{\bar{x}}{\sigma^2/N}+\frac{\mu_{prior}}{\tau_{prior}^2}\right], \left(\frac{1}{\sigma^2/N}+\frac{1}{\tau_{prior}^2}\right)^{-1}\right)$$

Similarly,

$$f(\sigma^2 \mid \mu, x) \propto (\sigma^2)^{-N/2-(1+\nu_{prior}/2)} e^{-\frac{1}{2\sigma^2}\left\{-N\left\{\mu^2-2\bar{x}\mu+\overline{x^2}\right\}+\nu_{prior}\sigma_{prior}^2\right\}}$$

From this, we can determine that,

$$\sigma^2 \mid \mu, x \sim Inv-\chi^2\left(\frac{\nu_{prior}\sigma_{prior}^2+N(\mu^2-2\bar{x}\mu+\overline{x^2})}{(\nu_{prior}+N)}, \nu_{prior}+N\right)$$

We can sample from the posterior distribution using the Gibbs Sampler in the following we,

(1) Sample $\mu_r \mid \sigma_{r-1}^2, x$ using direct methods

(2) Sample $\sigma_r^2 \mid \mu_r, x$ using direct methods

(3) Repeat

The, we can estimate $(\mu_0, {\sigma_0}^2)$ using $\hat{\mu} = \frac{1}{R}\sum_{r=R_1+1}^{R_1+R_2} \mu_r$ and $\hat{\sigma}^2 = \frac{1}{R}\sum_{r=R_1+1}^{R_1+R_2} \sigma_r^2$.

It is instructive to consider the large sample behavior of the iterates. Notice that, regardless of σ^2, we have, $\mu_r \to \mu_0$ as $N \to \infty$. In addition, $\sigma^2 \to \mu^2 - 2\overline{x}\mu + \overline{x^2} \to {\sigma_0}^2$ if $\mu \to \mu_0$. Hence, we expect this estimator to be consistent as well.

1.3.2 – The Linear Regression Model with Normal Errors

We can extend the logic of the previous sub-section to obtain a Bayesian estimator for the linear regression model with normal errors. Suppose that,

$$y_n = \beta' x_n + \sigma\varepsilon_n$$

and suppose that β and σ have the prior distributions,

$$\beta \sim N(\beta_{prior}, \Omega_{prior})$$

$$\sigma^2 \sim Inv - \chi^2(\sigma^2_{prior}, \nu_{prior})$$

Here, we will derive the posterior distribution for this model. We have,

$$f(\beta, \sigma^2 \mid y, x) \propto \sigma^{-N} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \beta' x_n)^2} e^{-\frac{1}{2}(\beta-\beta_{prior})'\Omega_{prior}^{-1}(\beta-\beta_{prior})} \frac{e^{-\nu_{prior}\sigma^2_{prior}/(2\sigma^2)}}{(\sigma^2)^{1+\nu_{prior}/2}}$$

Notice first that,

$$f(\beta \mid \sigma^2, y, x) \propto e^{-\frac{1}{2}\left[\beta'\left(\frac{1}{\sigma^2/N}\overline{xx'} + \Omega_{prior}^{-1}\right)\beta - 2\beta'\left(\frac{1}{\sigma^2/N}\overline{xy} + \Omega_{prior}^{-1}\beta_{prior}\right)\right]}$$

From this, we can determine that,

$$\beta \mid \sigma^2, y, x \sim N\left(\left[\tfrac{1}{\sigma^2/N}\overline{xx'}+\Omega_{prior}{}^{-1}\right]^{-1}\left[\tfrac{1}{\sigma^2/N}\overline{xy}+\Omega_{prior}{}^{-1}\beta_{prior}\right],\left[\tfrac{1}{\sigma^2/N}\overline{xx'}+\Omega_{prior}{}^{-1}\right]^{-1}\right)$$

Following a similar procedure used in 1.3.1, we can determine that,

$$\sigma^2 \mid \mu \sim Inv-\chi^2\left(\frac{\nu_{prior}\sigma^2_{prior}+\sum_{n=1}^{N}(y_n-\beta' x_n)^2}{(\nu_{prior}+N)},\nu_{prior}+N\right)$$

Once again, we are able to estimate this model via a combination of direct sampling and the Gibbs sampler.

From the above expressions, notice that,

$$(\overline{xx'}/\sigma^2+\tfrac{1}{N}\Omega_{prior}{}^{-1})^{-1}(\overline{xy}/\sigma^2+\tfrac{1}{N}\beta_{prior}\Omega_{prior}{}^{-1}) \xrightarrow{prob.} (\overline{xx'})^{-1}(\overline{xy})$$

so that the posterior mean converges to the OLS estimator. Also,

$$\tfrac{1}{N}(\overline{xx'}/\sigma^2+\tfrac{1}{N}\Omega_{prior}{}^{-1})^{-1}) \xrightarrow{prob.} \tfrac{1}{N}\sigma^2(\overline{xx'})^{-1}$$

which is the OLS estimate of the variance. Furthermore,

$$\frac{\nu_{prior}\sigma^2_{prior}+\sum_{n=1}^{N}(y_n-\beta' x_n)^2}{(\nu_{prior}+N)} \xrightarrow{prob.} \tfrac{1}{N}\sum_{n=1}^{N}(y_n-\beta' x_n)^2$$

which is the maximum likelihood estimator of the variance.

1.3.3 – The Binomial Probit Model

For the Binomial Probit model, the posterior distribution cannot be determined in closed form. The next best thing is to use the Gibbs sampler. Suppose that $y_n^* = \beta' x_n + \varepsilon_n$ where $\varepsilon_n \sim N(0,1)$ and $y_n = 1\{y_n^* \geq 0\}$. Suppose that the prior on β is given by $N(\beta_{prior}, B_{prior})$. We can write the posterior distribution as,

$$f(\beta \mid y,x) = \frac{f(\beta,y,x)}{f(y,x)} = \frac{f(y,x \mid \beta) f(\beta)}{f(y,x)}$$

Here, $f(y,x)$ is just a constant of integration, we can write,

$$f(\beta \mid y,x) \propto f(y,x \mid \beta) f(\beta)$$

We can determine that the posterior is given by,

$$f(\beta \mid x,y) \propto \frac{1}{\det(B_{prior})^{1/2}} e^{-\frac{1}{2}(\beta-\beta_{prior})' B_{prior}^{-1} (\beta-\beta_{prior})} \prod_{n=1}^{N} \Phi(-\beta' x_n)^{y_n} [1-\Phi(-\beta' x_n)]^{1-y_n}$$

The expression is clearly not a known density (and hence, we cannot easily sample from it). We could directly sample from this distribution (using the random walk sampler for example).

Alternatively, we could use data augmentation and apply the Gibbs sampler. Specifically, we would like to compute moments of the form,

$$\int_{\beta} h(\beta) f(x \mid \beta) d\beta$$

Let us introduce a variable, a, such that,

$$f(x \mid \beta) = \int_{a} f(x,a \mid \beta) da$$

We have,

$$\int_{\beta} h(\beta) f(x \mid \beta) d\beta = \int_{(\beta,a)} h(\beta) f(x,a \mid \beta) d(\beta,a)$$

This implies that we can compute $\int_{\beta} h(\beta) f(x \mid \beta) d\beta$ by sampling from $f(x,a \mid \beta)$. This approach will be effective of we can find an a such that $f(x,a \mid \beta)$ is easy to sample from. This process is called data-augmentation.

In this case, we will select $a = y^*$. We have,

$$f(\beta, y^* \mid y, x) \propto \left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_n^* - \beta' x_n)^2} 1\{y_n^* \geq 0\}^{y_n} 1\{y_n^* < 0\}^{1-y_n} \right)$$

$$* \frac{1}{(2\pi)^{J/2} \log \mid B_{prior} \mid} e^{-\frac{1}{2}(\beta - \mu_\beta)' B_{prior}^{-1} (\beta - \mu_\beta)}$$

We can write,

$$\beta \mid y^*, y, x \sim \left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_n^* - \beta' x_n)^2} \right) \frac{1}{(2\pi)^{J/2} \log |B_{prior}|} e^{-\frac{1}{2}(\beta - \mu_\beta)' B_{prior}^{-1} (\beta - \mu_\beta)}$$

$$\propto e^{-\frac{1}{2}\sum_{n=1}^{N}(y_n^* - \beta' x_n)^2 - \frac{1}{2}(\beta - \mu_\beta)' B_{prior}^{-1} (\beta - \mu_\beta)}$$

$$\propto e^{-\frac{1}{2}\beta' \left\{ \left[\sum_{n=1}^{N} x_n x_n' \right] + B_{prior}^{-1} \right\} \beta + \beta' \left\{ \sum_{n=1}^{N} x_n y_n^* + B_{prior}^{-1} \mu_\beta \right\}}$$

From this, it follows that,

$$\beta \mid y^*, y, x \sim N\left(\left[\sum_{n=1}^{N} x_n x_n' + B_{prior}^{-1} \right]^{-1} \left[\sum_{n=1}^{N} x_n y_n^* + B_{prior}^{-1} \mu_\beta \right], \left[\sum_{n=1}^{N} x_n x_n' + B_{prior}^{-1} \right]^{-1} \right)$$

In the case where a non-informative prior is used on β, we get,

$$\beta \mid y^*, x \sim N((x'x)^{-1} x' y^*, (x'x)^{-1})$$

Next, we have,

$$f(y_n^* \mid \beta, y_n = 1, x) \propto \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_n^* - \beta' x_n)^2} 1\{y_n^* \geq 0\}$$

$$f(y_n^* \mid \beta, y_n = 0, x) \propto \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_n^* - \beta' x_n)^2} 1\{y_n^* < 0\}$$

Hence, we have,

$$y_n^* \mid \beta, y_n = 1, x \sim TN(-\beta' x, 1, 0, \infty)$$

$$y_n^* \mid \beta, y_n = 0, x \sim TN(-\beta' x, 1, -\infty, 0)$$

The key thing to notice is that each conditional distribution can be easily sampled from (they are either normal or truncated normal). The Gibbs sampler then works by iterating these steps,

(i) Sample y_n^* conditional on β

(ii) Sample β conditional on y_n^*

1.3.4 – The Probit and Logit Models with Random Effects

Here, we consider two types of random effects models. We start with a binomial logit model with random effects. We assume that,

$$y_n^* = \beta' x_n + \varepsilon_n + \nu_{j[n]}$$

where $\varepsilon_n \sim EV(0,1)$ and $\nu_j \sim N(0,\sigma_\nu^2)$, where $j[n]$ denotes the group that observation n belongs to. We assume that we observe $y_n = 1\{y_n^* \geq 0\}$. Writing down the likelihood for y will require integrating over ν. To simply writing the likelihood, we employ data augmentation and we write down the augmented likelihood of (y,v). We have that,

$$f(y,v \mid x,\beta,\sigma_\nu) = \prod_{n=1}^{N}\left(\frac{e^{\beta' x_n+\nu_{j[n]}}}{1+e^{\beta' x_n+\nu_{j[n]}}}\right)^{y_n}\left(\frac{1}{1+e^{\beta' x_n+\nu_{j[n]}}}\right)^{1-y_n}\prod_{j=1}^{J}\frac{1}{\sigma_\nu\sqrt{2\pi}}e^{-\frac{1}{2}\nu_j^2/\sigma_\nu^2}$$

We can also specify priors as, $\beta_j \sim N(0,\sigma_{pr}^2)$ and $\sigma_\nu \sim U(0,a)$. We can write,

$$\pi(\beta,\sigma_\nu) = \left(\prod_{k=1}^{K}\frac{1}{\sigma_{pr}\sqrt{2\pi}}e^{-\frac{1}{2}\beta_k^2/\sigma_{pr}^2}\right)1\{0 \leq \sigma_\nu \leq a\}$$

Combining these, we can write the posterior as,

$$f(\beta,\sigma_\nu,v \mid y,x)$$

$$\propto \left[\prod_{n=1}^{N} \left(\frac{e^{\beta' x_n + \nu_{j[n]}}}{1+e^{\beta' x_n + \nu_{j[n]}}} \right)^{y_n} \left(\frac{1}{1+e^{\beta' x_n + \nu_{j[n]}}} \right)^{1-y_n} \right] \left(\prod_{j=1}^{J} \frac{1}{\sigma_\nu} e^{-\frac{1}{2}\nu_j^2/\sigma_\nu^2} \right) \left(\prod_{k=1}^{K} e^{-\frac{1}{2}\beta_k^2/\sigma_{pr}^2} \right) 1\{0 \le \sigma_\nu \le a\}$$

We could attempt to write a Gibbs sampler and see if any of the step can employ direct sampling. It is possible to employ direct sampling for σ_ν^2 if we change the prior to an inverse Gamma, but this does not help us that much. We would therefore mainly have to use techniques such as Slice Sampling, Adaptive Rejection Sampling, and Hamiltonian Monte Carlo to solve this problem.

Consider instead the binomial probit model with random effects. We assume that,

$$y_n^* = \beta' x_n + \varepsilon_n + \nu_{j[n]}$$

where $\varepsilon_n \sim N(0,1)$, $\nu_j \sim N(0,\sigma_\nu^2)$, and $y_n = 1\{y_n^* \ge 0\}$. Writing down the likelihood for y will again require integrating over ν. To simply writing the likelihood, we employ data augmentation, this time also augmenting the model with y^*. We have that the likelihood is,

$$f(y,\nu,y^* \mid \beta,\sigma_\nu^2,x) = \left(\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_n^* - \beta' x_n - \nu_{j[n]})^2} 1\{y_n^* \ge 0)^{y_n} 1\{y_n^* < 0)^{1-y_n} \right) \left(\prod_{j=1}^{J} \frac{1}{\sigma_\nu \sqrt{2\pi}} e^{-\frac{1}{2}\nu_j^2/\sigma_\nu^2} \right)$$

We employ the same prior for β_k as before, but we consider the inverse-Gamma prior for σ_ν^2. We have,

$$\pi(\beta,\sigma_\nu^2) = \left(\prod_{k=1}^{K} \frac{1}{\sigma_{pr}\sqrt{2\pi}} e^{-\frac{1}{2}\beta_k^2/\sigma_{pr}^2} \right) \frac{a^b}{\Gamma(a)} (\sigma_\nu^2)^{-a-1} e^{-b/\sigma_\nu^2}$$

Combining these, we can write the posterior as,

$$f(\beta,\sigma_\nu^2,\nu,y^* \mid y,x)$$

$$\propto \left(\prod_{n=1}^{N} e^{-\frac{1}{2}(y_n^* - \beta' x_n - v_{j[n]})^2} 1\{y_n^* \geq 0\}^{y_n} 1\{y_n^* < 0\}^{1-y_n} \right) \left(\prod_{j=1}^{J} \frac{1}{\sigma_v} e^{-\frac{1}{2} v_j^2 / \sigma_v^2} \right) \left(\prod_{k=1}^{K} e^{-\frac{1}{2}\beta_k^2 / \sigma_{pr}^2} \right) (\sigma_v^2)^{-a-1} e^{-b/\sigma_v^2}$$

We can determine that,

$$f(\beta \mid \sigma_v^2, v, y^*, y, x) \propto e^{-\frac{1}{2}\beta' \left[\sum_{n=1}^{N} x_n x_n' + (\sigma_{pr}^2 I)^{-1} \right] \beta - 2\beta' \sum_{n=1}^{N} (y_n^* - v_{j[n]}) x_n}$$

$$= N\left(\left[\sum_{n=1}^{N} x_n x_n' + (\sigma_{pr}^2 I)^{-1} \right]^{-1} \left[\sum_{n=1}^{N} (y_n^* - v_{j[n]}) x_n \right], \left[\sum_{n=1}^{N} x_n x_n' + (\sigma_{pr}^2 I)^{-1} \right]^{-1} \right)$$

$$f(\sigma_v^2 \mid \beta, v, y^* \mid y, x) \propto (\sigma_v^2)^{-J/2-a-1} e^{-\left[\frac{1}{2}\sum_{j=1}^{J} v_j^2 + b \right] / \sigma_v^2} = IG\left(a + J/2, b + \frac{1}{2}\sum_{j=1}^{J} v_j^2 \right)$$

$$f(v_j \mid \beta, \sigma_v^2, y^*, y, x) \propto e^{-\frac{1}{2}\left[\left(N_j + \frac{1}{\sigma_v^2} \right) v_j^2 - 2\left(\sum_{n=1}^{N} 1\{j[n] = j\}(y_n^* - \beta' x_n) \right) v_j \right]}$$

$$= N\left(\left(N_j + \frac{1}{\sigma_v^2} \right)^{-1} \sum_{n=1}^{N} 1\{j[n] = j\}(y_n^* - \beta' x_n), \left(N_j + \frac{1}{\sigma_v^2} \right)^{-1} \right)$$

$$f(y_n^* \mid \beta, \sigma_v^2, v, y = 1, x) \propto \prod_{n=1}^{N} e^{-\frac{1}{2}(y_n^* - \beta' x_n - v_{j[n]})^2} 1\{y_n^* \geq 0\} = TN(\beta' x_n + v_{j[n]}, 1, 0, \infty)$$

$$f(y_n^* \mid \beta, \sigma_v^2, v, y = 0, x) \propto \prod_{n=1}^{N} e^{-\frac{1}{2}(y_n^* - \beta' x_n - v_{j[n]})^2} 1\{y_n^* \leq 0\} = TN(\beta' x_n + v_{j[n]}, 1, -\infty, 0)$$

1.3.5 – Ideal Point Estimation

The Poole-Rosenthal estimator can be estimated using Bayesian methods. If normal priors are used, then the estimator can be implemented using the Gibbs sampler. Consider the model, $y_{n,t}^* = \delta_{t,0} + \delta_{t,1}\alpha_{n,1} + \delta_{t,2}\alpha_{n,2} + ... + \delta_{t,D}\alpha_{n,D} + \varepsilon_{n,t}$ where $y_{n,t} = 1\{y_{n,t}^* \geq 0\}$ and $\varepsilon_{n,t} \sim N(0,1)$.

We assume normal prior on α and δ given by $\alpha \sim N(\mu_\alpha, \Omega_\alpha)$ and $\delta \sim N(\mu_\delta, \Omega_\delta)$.

We implement the Gibbs sampler in three steps.

(i) Draw $y_{n,t}^* \mid y_{n,t}, \alpha_n, \delta_t$. We have,

$$y_{n,t}^* \mid y_{n,t} = 1, \alpha_n, \delta_t \sim TN(\delta_{t,0} + \delta_{t,1}\alpha_{n,1} + ... + \delta_{t,D}\alpha_{n,D}, 1, -\infty, 0)$$

$$y_{n,t}^* \mid y_{n,t} = 0, \alpha_n, \delta_t \sim TN(\delta_{t,0} + \delta_{t,1}\alpha_{n,1} + ... + \delta_{t,D}\alpha_{n,D}, 1, 0, \infty)$$

(ii) Draw $\delta_t \mid y_{n,t}^*, \alpha_n$.

Define $\tilde{\alpha}_n = (1, \alpha_n)$, then we have, $y_{n,t}^* = \delta_t \tilde{\alpha}_n + \varepsilon_{n,t}$. We can then show that,

$$\delta_t \mid y_{n,t}^*, \alpha_n \sim N((\alpha_n \alpha_n' + V_\alpha^{-1})^{-1}(\alpha_n y_{n,t}^* + V_\alpha^{-1} v_\alpha), (\alpha_n \alpha_n' + V_\alpha^{-1})^{-1})$$

(iii) Draw $\alpha_n \mid y_{n,t}^*, \delta_t$. In this case,

$$\alpha_n \mid y_{n,t}^*, \delta_t \sim N((\delta_t \delta_t' + V_\delta^{-1})^{-1}((\delta_{t,1}, ..., \delta_{t,D}) y_{n,t}^* + V_\delta^{-1} v_\delta), (\delta_t \delta_t' + V_\delta^{-1})^{-1})$$

1.3.6 – The Ordered Probit Model

Suppose that $y_n^* = \beta' x_n + \varepsilon_n$ where $\varepsilon_n \sim N(0,1)$. Suppose further that $y_n = j \Leftrightarrow \tau_{j-1} \le y_n^* \le \tau_j$. Let us normalize $\tau_0 = -\infty$, $\tau_1 = 0$, and $\tau_J = \infty$. We will

estimate this model using data augmentation. First, consider the likelihood of $(\beta, \tau, y^* \mid y)$. We have,

$$L(\beta, \tau, y^* \mid y) \propto \prod_{n=1}^{N} e^{-\frac{1}{2}(y_n^* - \beta' x_n)^2} \prod_{j=1}^{J} 1\{\tau_{j-1} \le y_n^* \le \tau_j\}^{1\{y_n = j\}}$$

From this, we can guess that the conditional distribution of β will be normal, the conditional distribution of y_n^* will be truncated normal, and the conditional distribution of τ will be uniform, provided that we choose appropriate prior distribution.[2]

Specifically, we select $\pi_\beta(\beta) = 1$ and $\pi_\tau(\tau) = 1\{\tau_1 = 0\} \prod_{j=2}^{J} 1\{\tau_{j-1} \le \tau_j\}$. We have that,

$$f(\beta, \tau, y^* \mid y) \propto 1\{\tau_1 = 0\} \prod_{j=2}^{J} 1\{\tau_{j-1} \le \tau_j\} \prod_{n=1}^{N} e^{-\frac{1}{2}(y_n^* - \beta' x_n)^2} \prod_{j=1}^{J} 1\{\tau_{j-1} \le y_n^* \le \tau_j\}^{1\{y_n = j\}}$$

We can determine that the conditional distributions are given by,

$$f(\beta \mid \tau, y^*, y) \sim N\left(\left[\frac{1}{N} \sum_{n=1}^{N} x_n x_n' \right]^{-1} \left[\frac{1}{N} \sum_{n=1}^{N} y_n^* x_n \right], \left[\frac{1}{N} \sum_{n=1}^{N} x_n x_n' \right]^{-1} \right)$$

$$f(y_n^* \mid \beta, \tau, y) \sim TN(\beta' x_n, 1, \tau_{y_n - 1}, \tau_{y_n})$$

$$f(\tau_k \mid \tau_1, ..., \tau_{k-1}, \tau_{k+1}, ..., \tau_J, \beta, y^* \mid y)$$

$$\sim U(\max\{\tau_{j-1}, \max\{y_n^* \mid y_n = j - 1\}\}, \min\{\tau_{j+1}, \min\{y_n^* \mid y_n = j\}\})$$

In practice, one may not want to impose the constraint of $\tau_1 = 0$ in the estimation, and instead post-process the Markov chain using, $\tau_{r,j}' = \tau_{r,j} - \tau_{r,1}$, for convergence reasons.

1.3.7 – The Multivariate Normal Model

[2] See Albert and Chib (1993).

The inverse Wishart distribution is given by,

$$f(\Omega \mid A, a) = \frac{(\det A)^{a/2} (\det \Omega)^{-(a+d+1)/2} e^{-tr(A\Omega^{-1})/2}}{2^{ad/2} \Gamma_d(m/2)}$$

where A is symmetric positive definite, d is the dimension of Ω and A, and $\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma(x + (1-j)/2)$ is the d-dimensional gamma function. It turns out the inverse Wishart is the conjugate prior distribution for the parameters covariance matrix of multivariate normal data.

If we assume that $X_n \sim N(\mu, \Omega)$ and we choose the prior distribution $\mu \sim N(b, B)$ and $\Omega \sim W^{-1}(A, a)$, we have the following posterior,

$$f(\mu, \Omega \mid X)$$

$$\propto \det(\Omega)^{-N/2} \left\{ \prod_{n=1}^{N} e^{-\frac{1}{2}(X_n - \mu)' \Omega^{-1} (X_n - \mu)} \right\} e^{-\frac{1}{2}(\mu - b)' B^{-1} (\mu - b)} (\det \Omega)^{-(a+d+1)/2} e^{-tr(A\Omega^{-1})/2}$$

We can derive the following conditional distributions.

$$\mu \mid \Omega \sim N\left(\left[N\Omega^{-1} + B^{-1} \right]^{-1} \left[\Omega^{-1} N\bar{X} + B^{-1} b \right], \left[N\Omega^{-1} + B^{-1} \right]^{-1} \right)$$

$$\Omega \mid \mu \propto W^{-1}\left(\sum_{n=1}^{N} (X_n - \mu)(X_n - \mu)' + A, N + a \right)$$

which means that we can estimate the parameters using the Gibbs sampler.

1.4 – Bayesian Hypothesis Testing

1.4.1 – Testing a Point Hypothesis against a Point Alternative

Consider the hypothesis test $H_0 : \theta = \theta_1$ against the alternative $H_A : \theta = \theta_2$. One could approach this problem by introducing a zero-one loss function. Specifically, we have the following,

$$L(\theta_1;\theta_1) = L(\theta_2;\theta_2) = 0, \qquad L(\theta_1;\theta_2) = L(\theta_2;\theta_1) = 1$$

Using the posterior, we can compute the expected loss of choosing θ_1 and θ_2,

$$E_\theta[L(\theta_1;\theta)] = \Pr(\theta_1 \mid x) * 0 + \Pr(\theta_2 \mid x) * 1 = \Pr(\theta_2 \mid x)$$

$$E_\theta[L(\theta_2;\theta)] = \Pr(\theta_1 \mid x) * 1 + \Pr(\theta_2 \mid x) * 0 = \Pr(\theta_1 \mid x)$$

This leads to the following decision rule,

$$d(x) = \begin{cases} \theta_1, & \frac{\Pr(\theta_2|x)}{\Pr(\theta_1|x)} < 1 \\ \theta_2, & \frac{\Pr(\theta_2|x)}{\Pr(\theta_1|x)} > 1 \\ \{\theta_1, \theta_2\}, & otherwise \end{cases}$$

The ratio $\frac{\Pr(\theta_2|x)}{\Pr(\theta_1|x)}$ is called the posterior odds.

More generally, we can consider the loss function,

$$L(\theta_1;\theta_1) = L(\theta_2;\theta_2) = 0, \qquad L(\theta_1;\theta_2) = 1, \qquad L(\theta_2;\theta_1) = a$$

We have that,

$$E_\theta[L(\theta_1;\theta)] = \Pr(\theta_1 \mid x) * 0 + \Pr(\theta_2 \mid x) * 1 = \Pr(\theta_2 \mid x)$$

$$E_\theta[L(\theta_2;\theta)] = \Pr(\theta_1 \mid x) * a + \Pr(\theta_2 \mid x) * 0 = a \Pr(\theta_1 \mid x)$$

The decision rule is now,

$$d(x) = \begin{cases} \theta_1, & \frac{\Pr(\theta_2|x)}{\Pr(\theta_1|x)} < a \\ \theta_2, & \frac{\Pr(\theta_2|x)}{\Pr(\theta_1|x)} > a \\ \{\theta_1, \theta_2\}, & otherwise \end{cases}$$

Let consider a specific example of this principle. We have two jars of coins, one jar of fair coins and one jar of biased coins that land on heads 60% of the time. Both jars fall on the floor and break and you now don't know whether a given coin is fair or biased. You know that a fraction α of the coins are fair and $1-\alpha$ are biased. You pick up a coin and flip it N times and find that it lands on heads Y times. You want to use the coin to cheat. Your loss is zero if you make the correct determination, 1 if you determine it is biased when it is not, and a if you determine that it is fair when it is not.

Notice that,

$$\Pr(Y \mid \theta_1) = \binom{N}{Y} \theta_1^{\,Y} (1-\theta_1)^{N-Y}, \qquad \Pr(Y \mid \theta_2) = \binom{N}{Y} \theta_2^{\,Y} (1-\theta_2)^{N-Y}$$

$$\Pr(\theta_1) = \alpha, \qquad \Pr(\theta_2) = 1-\alpha$$

We have that,

$$\Pr(\theta_1 \mid Y) = \frac{\Pr(Y \mid \theta_1)\Pr(\theta_1)}{\Pr(Y)} = \frac{\alpha \binom{N}{Y} \theta_1^{\,Y} (1-\theta_1)^{N-Y}}{\alpha \binom{N}{Y} \theta_1^{\,Y} (1-\theta_1)^{N-Y} + (1-\alpha) \binom{N}{Y} \theta_2^{\,Y} (1-\theta_2)^{N-Y}}$$

$$\Pr(\theta_2 \mid Y) = \frac{\Pr(Y \mid \theta_2)\Pr(\theta_2)}{\Pr(Y)} = \frac{(1-\alpha) \binom{N}{Y} \theta_2^{\,Y} (1-\theta_2)^{N-Y}}{\alpha \binom{N}{Y} \theta_1^{\,Y} (1-\theta_1)^{N-Y} + (1-\alpha) \binom{N}{Y} \theta_2^{\,Y} (1-\theta_2)^{N-Y}}$$

We can determine that the posterior odds is given by,

$$\underbrace{\frac{\Pr(\theta_1 \mid Y)}{\Pr(\theta_2 \mid Y)}}_{\textit{posterior odds}} = \underbrace{\frac{\Pr(Y \mid \theta_1)}{\Pr(Y \mid \theta_2)}}_{\textit{bayes factor}} \underbrace{\frac{\Pr(\theta_1)}{\Pr(\theta_2)}}_{\textit{prior odds}} = \underbrace{\frac{\binom{N}{Y} \theta_1^{\,Y} (1-\theta_1)^{N-Y}}{\binom{N}{Y} \theta_2^{\,Y} (1-\theta_2)^{N-Y}}}_{\textit{bayes factor}} \underbrace{\frac{\alpha}{1-\alpha}}_{\textit{prior odds}}$$

The decision rule depends on both the Bayes factor and the prior odds. If the prior odds are assumed to be equal, then a 0-1 loss function will lead one to select the model 1 when the Bayes factor is greater than 1. An asymmetric loss function will bias the decision rule in favor of one of the two alternatives. A different prior will have the same effect.

1.4.2 – Testing a Composite Hypothesis against a Composite Hypothesis

Consider instead the problem of testing the hypothesis $H_0 : \theta \in \Theta_1$ against the alternative $H_A : \theta \in \Theta_2$, where $\Theta_1 \cap \Theta_2 = \varnothing$. Consider the loss function,

$$L(d(x) = 1; \theta \in \Theta_1) = L(d(x) = 2; \theta \in \Theta_2) = 0,$$

$$L(d(x) = 1; \theta \in \Theta_2) = 1, \qquad L(d(x) = 2; \theta \in \Theta_1) = a$$

We can determine that the expected loss for $d(x) = 1$ and $d(x) = 2$ is given by,

$$E_\theta[L(d(x) = 1; \theta)] = \int_{\theta \in \Theta_1} f(\theta \mid X) d\theta * 0 + \int_{\theta \in \Theta_2} f(\theta \mid X) d\theta * 1 = \int_{\theta \in \Theta_2} f(\theta \mid X) d\theta$$

$$E_\theta[L(d(x) = 2; \theta)] = \int_{\theta \in \Theta_1} f(\theta \mid X) d\theta * a + \int_{\theta \in \Theta_2} f(\theta \mid X) d\theta * 0 = a \int_{\theta \in \Theta_1} f(\theta \mid X) d\theta$$

We can determine that the decision rule is,

$$d(x) = \begin{cases} \Theta_1, & \dfrac{\int_{\theta \in \Theta_2} f(\theta \mid X) d\theta}{\int_{\theta \in \Theta_1} f(\theta \mid X) d\theta} < a \\ \Theta_2, & \dfrac{\int_{\theta \in \Theta_2} f(\theta \mid X) d\theta}{\int_{\theta \in \Theta_1} f(\theta \mid X) d\theta} > a \\ \{\Theta_1, \Theta_2\}, & otherwise \end{cases}$$

Notice that if $\Theta_1 = \{\theta_1\}$ has a single value, then $\int_{\theta \in \Theta_1} f(\theta \mid X) d\theta = 0$ unless the prior puts a point mass on θ_1. A similar result holds for Θ_2. This implies that we cannot test a point

hypothesis using Bayesian techniques unless we put a point mass on the point hypothesis. This means that while one-sided tests are possible in a Bayesian framework, two-sided tests cannot be easily accommodated.

We can write,

$$\int_{\theta\in\Theta_1} f(\theta \mid X)d\theta = \int_{\theta\in\Theta_1} \frac{f(X \mid \theta)\pi(\theta)}{f(X)}d\theta = \frac{\int_{\theta\in\Theta_1} f(X \mid \theta)\pi(\theta)d\theta}{\int_{\theta\in\Theta_1\cup\Theta_2} f(X \mid \theta)\pi(\theta)d\theta}$$

$$\int_{\theta\in\Theta_2} f(\theta \mid X)d\theta = \int_{\theta\in\Theta_2} \frac{f(X \mid \theta)\pi(\theta)}{f(X)}d\theta = \frac{\int_{\theta\in\Theta_2} f(X \mid \theta)\pi(\theta)d\theta}{\int_{\theta\in\Theta_1\cup\Theta_2} f(X \mid \theta)\pi(\theta)d\theta}$$

The ratio is given by,

$$\underbrace{\frac{\int_{\theta\in\Theta_1} f(\theta \mid X)d\theta}{\int_{\theta\in\Theta_2} f(\theta \mid X)d\theta}}_{posterior\ odds} = \frac{\int_{\theta\in\Theta_1} f(X \mid \theta)\pi(\theta)d\theta}{\int_{\theta\in\Theta_2} f(X \mid \theta)\pi(\theta)d\theta} = \frac{\int_{\theta\in\Theta_1} f(X \mid \theta)\pi(\theta \mid \theta\in\Theta_1)\pi(\theta\in\Theta_1)d\theta}{\int_{\theta\in\Theta_2} f(X \mid \theta)\pi(\theta \mid \theta\in\Theta_2)\pi(\theta\in\Theta_2)d\theta}$$

$$= \underbrace{\frac{f(X \mid \theta\in\Theta_1)}{f(X \mid \theta\in\Theta_2)}}_{bayes\ factor} \underbrace{\frac{\pi(\theta\in\Theta_1)}{\pi(\theta\in\Theta_2)}}_{\substack{prior \\ odds}}$$

1.4.3 – Model Selection

Consider two models, 1 and 2. Let $M \in \{1,2\}$ in M_1 and M_2. For each of these two models, we can sample from the posterior distributions,

$$f(\theta_1 \mid X, M=1), \qquad f(\theta_1 \mid X, M=2)$$

However, to select a model, we need to determine the probability that each model is "correct" given the data,

$$\Pr(M=1\mid X)=\frac{\Pr(X\mid M=1)\Pr(M=1)}{\Pr(X)}$$

$$=\frac{\Pr(X\mid M=1)\Pr(M=1)}{\Pr(X\mid M=1)\Pr(M=1)+\Pr(X\mid M=2)\Pr(M=2)}$$

The choice of models will be determined by the posterior odds ratio,

$$\underbrace{\frac{\Pr(M=1\mid X)}{\Pr(M=2\mid X)}}_{posterier\ odds}=\underbrace{\frac{f(X\mid M=1)}{f(X\mid M=2)}}_{bayes\ factor}\underbrace{\frac{\Pr(M=1)}{\Pr(M=2)}}_{prior\ odds}$$

we have,

$$f(X\mid M=1)=\int_\theta f(X\mid\theta,M=1)d\theta$$

$$f(X\mid M=2)=\int_\theta f(X\mid\theta,M=2)d\theta$$

These quantities are, however, difficult to calculate.

1.4.4 – Computing Bayes Factors

Recall that,

$$f(\theta\mid X,M=j)=\frac{f(X\mid\theta,M=j)\pi(\theta\mid M=j)}{f(X\mid M=j)}$$

We have,

$$\log f(X\mid M=j)=\log f(X\mid\tilde{\theta},M=j)+\log\pi(\tilde{\theta}\mid M=j)-\log f(\tilde{\theta}\mid X,M=j)$$

for any given $\tilde{\theta}$ (e.g. the posterior mean). We can clearly calculate $f(X\mid\tilde{\theta},M=j)$ and $\pi(\tilde{\theta}\mid M=j)$. If we can calculate $\log f(\tilde{\theta}\mid X,M=j)$ as well, then we can calculate $f(X\mid M=j)$. Suppose that θ can be partitioned into two blocks $\theta=(\theta_1,\theta_2)$ such that $f(\theta_1\mid\theta_2,X,M=j)$ and $f(\theta_2\mid\theta_1,X,M=j)$ are known distributions. Notice that,

$$f(\theta \mid X, M = j) = f(\theta_2 \mid \theta_1, X, M = j) f(\theta_1 \mid X, M = j)$$

$$f(\theta_1 \mid X, M = j) = \int_{\theta_2} f(\theta_1, \theta_2 \mid X, M = j) d\theta_2 = \int_{\theta_2} f(\theta_1 \mid \theta_2, X, M = j) f(\theta_2 \mid X, M = j) d\theta_2$$

Let $\theta_2^{(r)}$ denote draws from the Gibbs sampler, so that $\theta_2^{(r)} \sim f(\theta_2 \mid X, M = j)$. We can estimate,

$$f(\tilde{\theta}_1 \mid X, M = j) \approx \frac{1}{R} \sum_{r=1}^{R} f(\tilde{\theta}_1 \mid \theta_2^{(r)}, X, M = j)$$

so that,

$$f(\tilde{\theta} \mid X, M = j) \approx f(\tilde{\theta}_2 \mid \tilde{\theta}_1, X, M = j) * \frac{1}{R} \sum_{r=1}^{R} f(\tilde{\theta}_1 \mid \theta_2^{(r)}, X, M = j)$$

This approach is due to Chib (1995).[3] This approach can be extended to more than two blocks and can be extend to cover Metropolis-Hastings sampling for the case where some of the full conditional distributions are not available in closed form (Chib and Jeliazkov, 2001).

A number of alternative techniques compute Bayes factors by jointly sampling over the parameter and model space. These include the methods of Carlin and Chib (1995), Dellaportas, Forster, and Ntzoufras (1998), and Green (1995). These methods and marginal likelihood methods have been evaluated by Cong, Han and Carling (2000). They determine that marginal likelihoods methods are superior in terms of ease of implementation and robustness, but that the Green's Reversible Jump MCMC often performs well.

1.5 – References

[3] See also Clarke (2000, 2001).

[1] Albert, J.H., and S. Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data". *Journal of the American Statistical Association*.

[2] Carlin, B.P., and Chib, S. (1995). "Bayesian Model Choice via Markov Chain Monte Carlo Methods". *Journal of the Royal Statistical Society, Series B* 57:473-484.

[3] Chib, S. (1995). "Marginal Likelihood from the Gibbs output". *Journal of the American Statistical Association* 90:1313-1321.

[4] Chib, Siddhartha, and Jeliazkov, Ivan (1999). "Marginal Likelihood from the Metropolis-Hastings Output". Working Paper.

[5] Clarke, Kevin A. (2000). "The Effect of Priors on Approximate Bayes Factors from MCMC Output". Working Paper.

[6] Clarke, Kevin A. (2001). "Testing Nonnested Models of International Relations: Reevaluating Realism". *American Journal of Political Science* 45:724-744.

[7] Dellaportas, P., Forster, J.J., and Ntzoufras, I. (1998). "On Bayesian Model and Variable Selection using MCMC". Working Paper.

[8] Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (1995). Bayesian Data Analysis.

[9] Gilks, W.R. and P. Wild. (1992). "Adaptive Rejection Sampling for Gibbs Sampling". *Applied Statistics* 41:337-348.

[10] Green, P.J. (1995). "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination". *Biometrika* 82:711-732.

[11] Han, Cong, and Bradley P. Carlin (2000). "MCMC Methods for Computing Bayes Factors: A Comparative Review". Working Paper.

[12] Jackman, Simon (2001). "Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference and Model Checking". *Political Analysis* 9:227-241.

[13] Jackman, Simon, Joshua Clinton, and Douglas Rivers (2004). "The Statistical Analysis of Roll Call Voting: A Unified Approach". American Political Science Review 98:355-370.

[14] Martin, Andrew, and Kevin Quinn (2002). "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999". *Political Analysis*

[15] McCulloch, R.E., N.G. Polson, and Peter E. Rossi (2000). "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters". *Journal of Econometrics*.

[16] McCulloch, R., and Peter E. Rossi (1994). "An Exact Likelihood Analysis of the Multinomial Probit Model". Journal of Econometrics.

[17] Press et. al. *Numerical Recipes in C*.

[18] Neal, Radford M. (2003). "Slice Sampling" *Annals of Statistics* 31:705-741.