Basics of Bayesian Econometrics

Notes for Summer School
Moscow State University, Faculty of Economics

Andrey Simonov[1]

June 2013

[1]University of Chicago, Booth School of Business. All errors and typos are of my own. Please report these as well as any other questions to asimonov@chicagobooth.edu.

Contents

Chapter 1

Introduction

1.1 Frequentists vs Bayesians

During the last decade the number of Bayesian papers in economics, finance and marketing has increased dramatically. However, frequentist inference still dominates economic literature, and Bayesian methods have had very little impact in economics so far. This is stunning given that a large proportion of statistical papers published today are explicitly Bayesian (Imbens and Wooldridge, 2007). What can drive this gap between proportion of Bayesian methods in statistics and economics? First, lets explain it by the historical reasons.

Being in the domain of economics, lets try to address this question with some economic intuition. Following Becker/Murphy approach, we can think of overall statistical knowledge in the field (i.e. economics, computer science, etc.) as a stock of specific human capital. A decade ago Bayesian methods were very restrictive due to lack of computational power necessary for simulations. With the development of technology these costs have become much lower, that is, the "price" of Bayesian inference has dropped. Assume for a moment that Bayesian methods are a perfect substitute to frequentist methods, and that all differences in methods come through some price of usage. Assume also that overall price of Bayesian inference today is lower than frequentist inference. If the stock of human capital would have the depreciation rate of one, economic theory tells us that researchers would use the methods with lower price, that is, switch to Bayesian inference. However, depreciation rate of human capital is usually much lower than one. This fact, coupled with key characteristic of human capital (it is inseparable from the owner, that is, cannon be sold), we can conclude that transition will take some time. The speed of this transition path is determined by the speed of depreciation of the human capital.

This analysis allows to conclude the following: even if the "price" of Bayesian inference nowadays is lower than "price" of frequentist inference, proportion of Bayesian papers would depend on the depreciation rate. It is intuitive that in statistics, the field where statistical knowledge is the core one, stock of statistical knowledge will be renewed much faster than in economics. This implies the difference in the extent to which Bayesian methods are used today.

Now lets relax the assumption that price of Bayesian inference is lower than price of

frequentists inference. Both have pros and cons, and mapping these into one dimension is problematic. There is an ongoing philosophical debate between frequentists and bayesians. Frequntists perspective on inference can be summarized as follows:

1. Parameters of interest are constants;
2. Estimation results in point estimates of these parameters and a confidence interval around it;
3. Estimation boils down to "accept" or "reject" conclusion about a given hypothesis.

Bayesians treat inference differently:

1. Parameters of interest are random variables;
2. Estimation results in a posterior distribution;
3. Posterior distribution contains all information about the parameter of interest.

The main philosophical difference is in treating the variance of an estimator. Frequentists use variance to accept or reject the hypothesis about the true value of θ. 95% confidence interval tells us that the true value of θ is inside of this interval with 95% probability. Bayesian probability intervals tell us that conditional on data and given a wide range or prior distributions, the posterior probability that θ would be in this interval is 95%. In a way frequentists through away information on the tails[1]. However, in practice Bayesian and frequentist inference is often very similar (that is, confidence intervals can be treated closely to the Bayesian probability intervals). The formal statement of this result is known as Bernstein-von Mises Theorem[2], which links Bayesian and frequentist inferences to the realm of asymptotics. This addresses another reasons often mentioned against using Bayesian inference: necessity to specify a prior. In fact, if prior distribution is not dogmatic[3], as the amount of data available increase prior distribution "washes out" of the posterior. Later I give examples where Bayesian and frequentist estimators are asymptotically the same.

Summing up, three main reasons are often mentioned against Bayesian methods:

- Its difficult to choose a prior distribution;
- Its necessary to specify a full parametric model;
- Its computationally complex.

We have just addressed the first reason using asymptotics equivalence argument. Second reason is a more serious one, but 1) a flexible specification can be chosen for the nuisance function, which would give robust results, and 2) Bayesian semi-parametric method have being developed. The final reason is more of a traditional one: with the development of MCMC methods Bayesian solutions sometimes are even less computationally burdensome (i.e. cases of bimodal underlying distributions and big parametric space).

[1] Calibrators go even further, using the point estimate as a true value.
[2] The theorem does not apply to irregular cases, i.e. time series with unit roots
[3] That is, does not put zero measure on any possible θ.

1.2 Road Map

In the next chapter I discuss foundations of Bayesian inference. It starts with the statement of Bayes' Theorem and its main ingredients: prior and posterior distributions, likelihood, as well as predictive distribution and point estimator. In section 2.2 I discuss two fundamental principles: likelihood and sufficiency, both obeyed by Bayesian methods. In section 2.3 I define conjugacy of distributions and give several examples. Section 2.4 shows how simple linear regressions is done in Bayesian context.

In chapter 3 I discuss numerical methods that complement bayesian analysis. Section 3.1 contains definition of Monte Carlo simulation, and section 3.2 discuss basic Markov chain theory. Section 3.3 combines the two and defines Monte Carlo Markov Chain method. Section 3.4 contains a number of popular algorithms for the simulation of posterior, including Gibbs sampler and Metropolis-Hastings algorithm. Section 3.5 has some applications.

1.3 Attributions and Literature

I used three main sources for reference. Some examples are taken from Imbens and Wooldridge (2007) [3]. Discussion of algorithms is based on Rossi, Allenby and McCulloch (2005) [5]. For theoretical results are taken from Robert (2007) [4]. Additionally, I used notes of L.P. Hansen (2013) [2] as a reference for stochastic processes and ergodic theory. Of course, all errors and typos are of my own.

1.4 Things out of scope

There is a lot of nice developments and applications of Bayesian methods, as well as things that go together with Bayesian methods. Unfortunately, we will not be able to talk about all of these. These notes do not discuss (for various reasons):

- Bayesian decision theory;
- Testing;
- Numerical integration methods;
- Optimization methods;
- Kalman Filter and HMM models;
- Model Averaging.
- and a lot of other things.

Chapter 2

Foundations

2.1 Bayes' Theorem and its Ingredients

2.1.1 Bayes' Theorem

We start the discussion with the definition of the Bayes' Theorem. Let A and B be two events, $P(B) \neq 0$, where $P(\cdot)$ denotes probability. Then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)} = \frac{P(B|A)P(A)}{P(B)}$$

Notice that this implies that $P(A|B)P(B) = P(B|A)P(A)$, which is intuitive as both of these are equal to the joint probability $P(A, B)$. In the continuous case this can be written as

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \tag{2.1}$$

where

- θ is the parameter of interest and x is data;
- $\pi(\theta)$ is a marginal distribution of θ;
- $f(x|\theta)$ is a conditional distribution of x given θ;
- $\int f(x|\theta)\pi(\theta)d\theta$ is a marginal distribution of x;
- $\pi(\theta|x)$ is a conditional distribution of θ given x.

$\pi(\theta)$ is what is also called a *prior distribution*, $f(x|\theta)$ is called *likelihood* of θ given x, and $\pi(\theta|x)$ is called *posterior distribution*. Sometimes we will refer to the likelihood as $l(\theta|x)$ (which is exactly $f(x|\theta)$: $l(\theta|x) = f(x|\theta)$), just to emphasize that likelihood is a function of θ.

This is the stage where Bayesian and frequentists inference diverge. Frequentists use the likelihood $f(x|\theta)$ to find a θ which maximizes it (that is, find MLE estimator). Bayesians

are interested in the posterior distribution, which is the probability of observing a particular θ *given* the data.

Computation of the posterior distribution can seem quite burdensome, especially due to integration in the denominator. Indeed, updating the posterior would imply integration in every step. However, expression in the denominator of (2.1) does not depend on θ, which implies that posterior distribution is proportional to the likelihood times a prior distribution:

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$$

Knowing likelihood and prior distribution, it is sometimes easier to scale their product by some constant so that it integrates to one over θ rather than compute the marginal of x. In particular, it allows to update posterior several times before this normalization (so it will be done only once instead of being done in each stage). Moreover, some numerical methods do not require knowing the exact posterior distribution.

2.1.2 Predictive Distribution

Likelihood, prior and posterior are the main ingredients of any Bayesian problem. The only other distribution that sometimes appear is called *predictive distribution*:

$$g(x^{T+1}|x) = \int g(x^{T+1}|\theta, x)\pi(\theta|x)d\theta$$

That is, the conditional distribution of the future observation given all observations up to date is a function only of current data, not of θ (θ is integrated out). This contrasts the frequentists method predicting the future using the point estimator.

2.1.3 Point estimator

Finding a point estimator in a Bayesian framework contradicts the fundamental ideas of Bayesian estimation. However, one can easily compute the posterior expectation of function of interest $h(\theta)$:

$$E_{\theta|x}(h(\theta) = \int h(\theta)\pi(\theta|x)d\theta = \frac{\int h(\theta)f(x|\theta)\pi(\theta)d\theta}{\int f(x|\theta)\pi(\theta)d\theta}$$

Another way to get to a point estimate is to look at the mode of θ and find $h(\theta_{mode})$.

2.2 Two Fundamental Principles

Bayesian paradigm is highly accepted in the statistical literature as it follows two fundamental principles: Likelihood Principle and Sufficiency Principle.

2.2.1 Sufficiency Principle

Definition 2.2.1. *When $x \sim f(x|\theta)$, a function T of x (also called a statistics) is said to be sufficient if the distribution of x conditional upon $T(x)$ does not depend on θ:*

$$f(x|T(x), \theta) = f(x|T(x)) \tag{2.2}$$

This says that given $T(x)$ conditioning on θ does not provide us with any extra information about x. Under some regularity conditions (see Lehmann and Casella (1998)) Fisher-Neyman factorization theorem holds:

$$f(x|\theta) = g(T(x)|\theta)h(x|T(x)) \tag{2.3}$$

where $g(\cdot)$ is a density of $T(x)$. Intuitively, this implies that we can find a mediating function $T(x)$ such that θ affects x only through this function[1]. We can define a minimal sufficient statistics as a function of all other sufficient statistics. I.e. if $X_1, \ldots, X_n$ are independent draws from Bernoulli distribution with probability p, $\sum_{i=1}^{N} X_i$ is a sufficient statistic for p, and if $X_1, \ldots, X_n$ are independent draws from Normal distribution with mean μ and known variance σ^2, sufficient statistics is $\frac{1}{n}\sum_{i=1}^{N} X_i = \bar{X}$. The following is the *sufficiency principle* developed by Fisher:

Definition 2.2.2. *Two observations x and y factorizing through the same value of a sufficient statistic T, that is, $T(x) = T(y)$, must lead to the same inference on θ.*

In general, sufficiency is a very powerful concept in case of exponential distributions (as minimal sufficient statistics is often two-dimensional), but can be not so useful in other cases (for a number of distributions order statistics is minimal sufficient, and it has the same dimension as the original sample).

2.2.2 Likelihood Principle

Definition 2.2.3. *The information[2] brought by an observation x about θ is entirely contained in the likelihood function $l(\theta|x) = f(x|\theta)$. Moreover, if x_1 and x_2 are two observations depending on the same parameter θ, such that there exist a constant c satisfying*

$$l_1(\theta|x_1) = cl_2(\theta|x_2)$$

for every θ, they then bring the same information about θ and must lead to identical inferences.

Likelihood principle is only valid when inference is made about the same parameter θ and if θ includes every unknown factor of the model. The following example was taken from Robert (2007):

[1] Formally, refer to Rao-Blackwell theorem, which says that in case of estimators under a convex loss function, optimal procedures would only depend on sufficient statistics (see Robert (2007), Chapter 2).

[2] Information in this case refer to any kind of information that can help to make inference on θ, not information in Fisher information sense

Example 2.2.1. *While working on the audience share of a TV series,* $0 \leq \theta \leq 1$ *representing the part of the TV audience, an investigator found nine viewers and three non-viewers. If no additional information is available on the experiment, two probability models at least can be proposed:*

- *the investigator questioned 12 person, thus observed* $x \sim B(12, \theta)$ *with* $x = 9$ *(B is Binomial distribution);*
- *the investigator questioned N person until she obtained 3 non-viewers, with* $N \sim Neg(3, 1-\theta)$ *and* $N = 12$ *(Neg is negative binomial distribution).*

That is, the random quantity can either be 9 or 12. Importantly, in both models likelihood is proportional to

$$\theta^3(1-\theta)^9$$

so likelihood principle implies that inference on θ should be the same.

Bayesian approach is based on the posterior distribution, which depends on x only through the likelihood $l(\theta|x)$, and thus automatically satisfy the Likelihood Principle.

2.3 Conjugate distributions

Bayesian inference is based on the posterior distribution; but for a general likelihood and prior distribution, a nice analytical expression for a posterior is not guaranteed. Nowadays this is not a restriction due to powerful simulation methods; however, having a nice expression for a posterior would simplify inference.

One way to deal with this is to require prior distribution to be *conjugate* to the likelihood:

Definition 2.3.1. *A prior is said to be conjugate to the likelihood if the posterior derived from the prior and likelihood is in the same class of distributions as the posterior.*

Due to the properties of the exponent function exponential family of distributions have a number of conjugate priors.

2.3.1 Example: Normal and Normal

Example 2.3.1. *Suppose a conditional distribution of X given* μ *is* $N(\mu, 1)$*. Also suppose that prior distribution of* θ *is* $N(0, 100)$*. Assume we observe a single observation x. What is the posterior distribution* $f(\theta|x)$*?*

Computing the posterior:

$$\begin{aligned} f(\theta|x) &\propto \exp(-\frac{1}{2}(x-\mu)^2)\exp(-\frac{1}{2}\frac{\mu^2}{100}) = \\ &= \exp -\frac{1}{2}(x^2 - 2x\mu + \mu^2 + \mu^2/100) \propto \\ &\propto \exp(-\frac{(\mu - 100x/101)^2}{2(100/101)}) \end{aligned}$$

That is, posterior distribution $f(\theta|x)$ is $N(100x/101, 100/101)$. We can conclude that two normal distributions are conjugate, so normal prior times normal likelihood gives a normal posterior.

Notice that variance of likelihood distribution was much less than variance of the prior distribution, and the resulting posterior is closer to the likelihood than to the prior. One can intuitively expect that posterior efficiently weight the information given by prior distribution and likelihood, and the weights are determined by the variance. Lets examine a more general case to see if it holds.

Example 2.3.2. *Suppose a conditional distribution of X given μ is $N(\mu, \sigma^2)$ (σ^2 is known). Also suppose that prior distribution of θ is $N(\mu_0, \tau^2)$. Assume we observe a single observation x. What is the posterior distribution $f(\theta|x)$?*

$$\begin{aligned} f(\theta|x) &\propto \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)\exp(-\frac{1}{2\tau^2}(\mu-\mu_0)^2) = \\ &= \exp -\frac{1}{2}(\frac{x^2}{\sigma^2} - \frac{2x\mu}{\sigma^2} + \frac{\mu^2}{\sigma^2} + \frac{\mu^2}{\tau^2} - \frac{2\mu\mu_0}{\tau^2} + \frac{\mu_0^2}{\tau^2}) \propto \\ &\propto \exp(-\frac{(\mu - \frac{x/\sigma^2+\mu_0/\tau^2}{1/\sigma^2+1/\tau^2})^2}{2\frac{1}{1/\tau^2+1/\sigma^2}}) \end{aligned}$$

so posterior distribution of $f(\theta|x)$ is $N(\frac{x/\sigma^2+\mu_0/\tau^2}{1/\sigma^2+1/\tau^2}, \frac{1}{1/\tau^2+1/\sigma^2})$. As expected, posterior mean is a weighed average of the prior mean μ_0 and of the observation x, and weights are proportional to the precision (1 over the variance) $1/\sigma^2$ and $1/\tau^2$. Posterior precision is just the sum of two precision components.

Intuitively variance of the prior distribution reflect to what extent researcher is sure about his prior information. In case of τ^2 being a big number researcher puts less weight on the prior and more weight on the observation. Notice the researcher can even set $\tau^2 = \infty$ (which is not a proper distribution anymore). However, posterior will be perfectly well defined:

$$f(\theta|x) \sim N(x, \sigma^2)$$

Setting variance of the prior distribution to the infinity is the same as saying that we know nothing about μ a priori. In general, we would like to find an uninformative prior which would allow to identify the posterior distribution but put as few weight on itself as possible.

To illustrate the idea of sufficiency principle lets look at another closely related example.

Example 2.3.3. *Again, suppose a conditional distribution of X given μ is $N(\mu, \sigma^2)$ (σ^2 is known). Also suppose that prior distribution of θ is $N(\mu_0, \tau^2)$. Assume we observe N independent draws $x_1, \ldots, x_N$. What is the posterior distribution $f(\theta|x_1, \ldots, x_N)$?*

$$f(\theta|x_1,\ldots,x_N) \propto \exp(-\frac{1}{2\tau^2}(\mu-\mu_0)^2)\prod_{i=1}^{N}\exp(-\frac{1}{2\sigma^2}(x_i-\mu)^2) \propto$$
$$\propto \exp(-\frac{(\mu-\frac{\sum_i x_i/\sigma^2+\mu_0/\tau^2}{N/\sigma^2+1/\tau^2})^2}{2\frac{1}{1/\tau^2+N/\sigma^2}})$$

so posterior distribution of $f(\theta|x_1,\ldots,x_N)$ is $N(\frac{\sum_i x_i/\sigma^2+\mu_0/\tau^2}{N/\sigma^2+1/\tau^2}, \frac{1}{1/\tau^2+N/\sigma^2})$. Posterior distribution $f(\theta|x_1,\ldots,x_N)$ depends only on sufficient statistic $\sum_i x_i$.

2.3.2 Example: Binomial and Beta

For an example without normal distribution consider the following:

Example 2.3.4. *Conditional distribution of X is $B(\theta)$, where B is Bernoulli distribution, and prior distribution is $Beta(\alpha,\beta)$. Assume we observe N independent draws $x_1,\ldots,x_N$. What is the posterior distribution $f(\theta|x_1,\ldots,x_N)$?*

In this case joint density of the data is

$$f(x_1,\ldots,x_N|\theta) = \theta^{\sum_i x_i}(1-\theta)^{N-\sum_i x_i}$$

and prior is proportional to

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Then

$$f(\theta|x_1,\ldots,x_N) \propto \theta^{\alpha-1+\sum_i x_i}(1-\theta)^{\beta-1+N-\sum_i x_i}$$

That is posterior distribution is $Beta(\alpha+\sum_i x_i, \beta+N-\sum_i x_i)$

2.3.3 Asymptotic Equivalence

Lets go back to the example with N normal observations. What happens as $N\to\infty$?

We can rewrite

$$f(\theta|x_1,\ldots,x_N) \sim N(\frac{\sum_i x_i/\sigma^2+\mu_0/\tau^2}{N/\sigma^2+1/\tau^2}, \frac{1}{1/\tau^2+N/\sigma^2})$$

as

$$f(\theta|x_1,\ldots,x_N) \sim N(\frac{\sum_i x_i/\sigma^2}{N/\sigma^2+1/\tau^2}+\frac{\mu_0/\tau^2}{N/\sigma^2+1/\tau^2}, \frac{1}{1/\tau^2+N/\sigma^2})$$

For $N\to\infty$ $limE(\theta|x)=\bar{x}$ and $limV(\theta|x)=0$. Examine the distribution of $\sqrt{N}(\mu-\bar{x})$:

$$\sqrt{N}(\bar{x}-\mu)|x_1,\ldots,x_N \sim N(0,\sigma^2)$$

We can conclude that as N increases 1) effect of prior distribution disappears, as notes before; 2) large sample properties of the estimator are the same as in a frequentists analysis.

In general this result is known as Bernstein-von Mises Theorem, and is applicable the a wide range of distributions (i.e. as an exercise do the same for Binomial and Beta from the previous section). See Imbens and Wooldridge (2007) for a statement of the Theorem.

2.4 Bayesian Regressions

2.4.1 Multiple Regression

Lets give several example of practical econometrics problems. Suppose we want to analyze simple multiple regression:

$$y_i = x_i'\beta + \epsilon_i \quad \epsilon_i \sim iidN(0, \sigma^2)$$

that is

$$y \sim N(X\beta, \sigma^2 I)$$

This is a conditional distribution of y given x. Assume that distribution of x depends on some other parameter ψ, and ψ is a priori independent of β and σ^2. We can write the posterior of ψ, β, σ^2 as

$$\pi(\psi, \beta, \sigma^2|y, X) \propto (\pi(\psi)f(X|\psi))(\pi(\beta, \sigma^2)f(y|X, \beta, \sigma^2) \tag{2.4}$$

We can focus on the second term of (2.4) which does not depend on ψ. Now we need to specify a prior $\pi(\beta, \sigma)$. We need to find a natural conjugate prior for

$$f(y|X, \beta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)) \tag{2.5}$$

Natural conjugate prior for σ^2 and β would be proportional to (2.5). Notice that exponent term has quadratic expression for β in it. This allows to assume that normal prior for β would be conjugate. Lets rewrite the term in the exponent as a usual quadratic expression in β by projecting y on X and taking a part independent of β out:

$$f(y|X, \beta, \sigma^2) \propto (\sigma^2)^{-n/2} \exp(-\frac{y'M_x y}{2\sigma^2}) \exp(-\frac{1}{2\sigma^2}(\beta - \beta_{OLS})'(X'X)(\beta - \beta_{OLS})) \tag{2.6}$$

where M_x is a projection on space orthogonal to X: $M_x = I - X(X'X)^{-1}X$. The first term in (2.6) suggests inverse gamma prior of σ^2, while the term in the exponent suggest normal prior for β. Thus, we have split the prior into two parts:

$$\pi(\beta, \sigma) = \pi(\sigma^2)\pi(\beta|\sigma^2)$$

$$\pi(\sigma^2) \propto (\sigma^2)^{-\nu_0/2+1} \exp(-\frac{\nu_0 s_0^2}{2\sigma^2})$$

$$p(\beta|\sigma^2) \propto (\sigma^2)^{-k/2} \exp\{-\frac{1}{2\sigma^2}(\beta - \bar{\beta})'A(\beta - \bar{\beta})\}$$

where $\pi(\sigma^2)$ is standard inverse gamma with $\alpha = \nu_0/2$ and $\beta = \frac{\nu_0 s_0^2}{2}$ and $p(\beta|\sigma^2) \sim N(\bar{\beta}, \sigma^2 A^{-1})$.

Now we can express the posterior $\pi(\beta, \sigma^2|y, X)$. Given the conjugacy posterior will be the same form as the prior:

$$\pi(\beta, \sigma^2|y, X) \propto (\sigma^2)^{-(n+\nu_0)/2+1} \exp(-\frac{(\nu_0 s_0^2 + ns^2)}{2\sigma^2}) \times \tag{2.7}$$

$$\times (\sigma^2)^{-k/2} \exp\{-\frac{1}{2\sigma^2}(\beta - \tilde{\beta})'(X'X + A)(\beta - \tilde{\beta})\} \tag{2.8}$$

where

$$\tilde{\beta} = (X'X + A)^{-1}(X'X\beta_{OLS} + A\bar{\beta})$$
$$ns^2 = (y - X\tilde{\beta})'(y - X\tilde{\beta}) + (\tilde{\beta} - \bar{\beta})'A(\tilde{\beta} - \bar{\beta})$$

This implies that

$$\beta|\sigma^2, y, X \sim N(\tilde{\beta}, \sigma^2(X'X + A)^{-1})$$
$$\sigma^2|y, X \sim IG((\nu_0 + n)/2, (\nu_0 s^2 + ns^2)/2)$$

The Bayes estimator of the posterior mean is $\tilde{\beta}$. Notice that $\tilde{\beta}$ is a weighted average of the prior mean and the OLS estimator, β_{OLS}. Also, notice that if precision parameter A goes to zero, $\tilde{\beta} = \beta_{OLS}$, and $ns^2 = \hat{e}'\hat{e}$. Other way of thinking about it is that as we accumulate for observations, weight of prior information goes to zero, and $\tilde{\beta}$ converge to β_{OLS}.

2.4.2 Multivariate Regression

Another example of regression which is more general but where conjugate priors can still be found is multivariate regression. We have a set of equations:

$$\begin{aligned} y_1 &= X\beta_1 + \epsilon_1 \\ &\vdots \\ y_m &= X\beta_m + \epsilon_m \end{aligned} \tag{2.9}$$

Each equation has a set of n observations, with X being the same and ϵ being correlated across regressions (but iid across observations!). That is, one observation (across equations) can be written as

$$y_r = B'x_r + \epsilon_r \quad \epsilon \sim iid\ N(0, \Sigma)$$

Σ is of dimension m, and columns of B correspond to equations $1, \ldots, m$ (B is $(k \times m)$). We would like to know the posterior of B and Σ. Denote $\boldsymbol{\beta} = vec(B)$. Without going into details (please see Rossi et al. (2005), Section 2.8.5) we can say that this problem has natural conjugate priors:

$$p(\Sigma, B) = p(\Sigma)p(B|\Sigma)$$
$$\Sigma \sim IW(\nu_0, V_0)$$
$$\boldsymbol{\beta}|\Sigma \sim N(\bar{\boldsymbol{\beta}}, \Sigma \otimes A^{-1})$$

Posterior distribution is

$$\Sigma|Y, X \sim IW(\nu_0 + n, V_0 + S)$$
$$\boldsymbol{\beta}|\Sigma, Y, X \sim N(\tilde{\boldsymbol{\beta}}, \Sigma \otimes (X'X + A)^{-1})$$
$$\tilde{\boldsymbol{\beta}} = vec(\tilde{B}) \qquad \tilde{B} = (X'X + A)^{-1}(X'XB_{OLS} + A\bar{B})$$
$$S = (Y - X\tilde{B})'(Y - X\tilde{B}) + (\tilde{B} - \bar{B})'A(\tilde{B} - \bar{B})$$

2.4.3 Seemingly Unrelated Regression

Lets generalize standard regression even further. Consider (2.9) having different X in each set of regressions $(1, \ldots, m)$. That is, regressions are related only by the correlation in ϵ:

$$y = X\beta + \epsilon$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \qquad X = \begin{bmatrix} X_1 & 0 & 0 & 0 \\ 0 & X_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & X_m \end{bmatrix} \qquad \epsilon \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{bmatrix} \tag{2.10}$$

$$var(\epsilon) = \Sigma \otimes I_n$$

There are no conjugate priors for this problem: assuming normal prior for $\beta|\Sigma$ and inverted Wishart for Σ, we cannot integrate out Σ from the conditional distribution for β. it turns out that this problem can easily be solved via MCMC methods (Gibbs sampler in particular), which we discuss in the next section.

Chapter 3

MCMC Methods

Although it is nice to have a conjugate prior distribution, it is somewhat restrictive. As mentioned before, current developments of computational power for numerical methods allow to tackle posterior distributions which are much more general. This section discuss *Markov Chain Monte Carlo (MCMC)* methods that are available for simulating the posterior distribution.

3.1 Monte Carlo simulation

Say we need to approximate an integral of the form

$$\int_{\Theta} g(\theta) f(x|\theta) \pi(\theta) d\theta \tag{3.1}$$

One way to go is to use numerical integration, i.e. polynomial quadratures for cases of distribution close to normal (see Robert 2007, Section 6.2.1, for a short description and references). Numerical integration is a vast topic which we would not be able to cover here.

We can go in another direction and take advantage of the fact that $\pi(\theta)$ is a known density. If it is possible to sample from this density we can generate m draws $\theta_1, \ldots, \theta_m$ and compute

$$\frac{1}{m} \sum_{i=1}^{m} g(\theta_i) f(x|\theta_i)$$

From the LLN we know that this (almost surely) converge to the expectation of $g(\theta)$

$$\frac{1}{m} \sum_{i=1}^{m} g(\theta_i) f(x|\theta_i) \xrightarrow{a.s.} \int_{\Theta} g(\theta) f(x|\theta) \pi(\theta) d\theta$$

as desired. Similarly, we can also sample for the posterior $\pi(\theta|x)$:

$$\frac{1}{m} \sum_{i=1}^{m} g(\theta_i) \xrightarrow{a.s.} \frac{\int_{\Theta} g(\theta) f(x|\theta) \pi(\theta) d\theta}{\int_{\Theta} f(x|\theta) \pi(\theta) d\theta}$$

3.1.1 Importance sampling

Monte Carlo methods can be generalized even further. Sampling from $\pi(\cdot)$ or $\pi(\cdot|x)$ can be quite complicated. But luckily we are not restricted to the simulation only from these two: if $h(\cdot)$ is a probability density such that $supp(h)$ includes support of $g(\theta)f(x|\theta)\pi(\theta)$, integral in (3.1) can be written as

$$\int_\Theta \frac{g(\theta)f(x|\theta)\pi(\theta)}{h(\theta)}h(\theta)d\theta \tag{3.2}$$

Function $h(\cdot)$ is importance sampling function: generating $\theta_1, \ldots, \theta_k$ from h we can approximate (3.2) by

$$\frac{1}{k}\sum_{i=1}^{k} g(\theta_i)w_i(\theta_i)$$

where $w_i(\theta_i) = \frac{f(x|\theta_i)\pi(\theta_i)}{h(\theta_i)}$, and again by LLN

$$\frac{1}{k}\sum_{i=1}^{k} g(\theta_i)w_i(\theta_i) \xrightarrow{a.s.} \int_\Theta g(\theta)f(x|\theta)\pi(\theta)d\theta$$

Approximation of the expectation of $g(\theta)$ under $\pi(\cdot)$ can be computed as

$$\mathbb{E}^\pi(g(\theta)|x) \sim \frac{\sum_{i=1}^k g(\theta_i)w(\theta_i)}{\sum_{i=1}^k w(\theta_i)} \tag{3.3}$$

This approximation does not depend on the normalizing constants in $h(\theta)$, $f(x|\theta)$ and $\pi(\theta)$, which allows to use it in the setting with unknown constants.

Importance sampling is a very powerful tool; however, there are two important caveats. First, support of $h(\cdot)$ should include support of $g(\theta)f(x|\theta)\pi(\theta)$ as said above. This is not so restrictive. Second, although (3.3) theoretically converges to $\mathbb{E}^\pi(g(\theta)|x)$, choice of function h determines the variance of the estimator (3.3). If h is chosen such that it is far from $g(\theta)\pi(\theta|x)$, $\mathbb{E}^h(g^2(\theta)w^2(\theta))$ would be not finite, and variance of (3.3) would be infinity.

Bottom line: choose $h(\cdot)$ such that

- It is easy to simulate from it;
- Support of $h(\theta)$ covers the support of $g(\theta)f(x|\theta)\pi(\theta)$;
- $h(\theta)$ is as close to $g(\theta)f(\theta|x)$ as possible.

From which $h(\cdot)$ it is easy to sample? There is a number of distributions for which efficient methods of simulation are available. We discuss the most common ones in the next subsection.

3.1.2 Methods of simulations: frequently used distributions

Simulation seems to be an easy solution for any kind of problem: if we know the posterior why not just simulate from it. However, in generating random numbers for an arbitrary

(and possibly high-dimensional) distribution has no general purpose solution. Later in Chapter we will discuss methods that exploit special structure of Bayesian models, but the basic building block for these methods is efficient simulation from the simple frequently used distributions. I discuss several examples here; interested readers can refer Rossi et al (2005), Section 2.11.

All methods for continuous random generation start with *uniform* pseudo-random number generator. From uniform randoms one can get to *normal, gamma and chi-squared random variates* i.e. via inverse cdf method. Inverse cdf method does exactly what one would expect: it takes a uniform draw and numerically computes random variable that corresponds to this quantile level.

Inverse cdf methods can also be applied to sample from *truncated distributions*. I.e. if we want to simulate normal draws conditional on being greater than zero (think about Tobit models), we can invert the following cdf:

$$G_X(x) = \frac{F(x) - F(0)}{F(\infty) - F(0)}$$

Solving this for x:

$$x = F^{-1}(G_X(x)(F(\infty) - F(0)) + F(0))$$

and we can sample it by sampling $U(0,1)$ for $G_X(x)$.

Sampling *multivariate normals* is easily done with the help of Cholesky matrix: we can sample standard normals $z_1, \ldots, z_k$, and then compute

$$x = U'z + \mu \sim N(\mu, \Sigma)$$

where $\Sigma = U'U$.

Rossi et al. (2005) discuss how to sample from Student t, Wishart and Inverted Wishart, Multinomial and Dirichlet distributions via other simple algebraic procedures.

3.2 Basic Markov Chain Theory

In this section we discuss the basics of Markov chains.

Definition 3.2.1. *Markov chain is a sequence of random variables* $X_i \;\; i = 1, 2, \ldots$ *such that*

$$P(X_{n+1} = x | X_1, \ldots, X_n) = P(X_{n+1} = x | X_n) \tag{3.4}$$

(3.4) is also referred to as Markov property.

Markov chains can be both discrete and continuous. In order for the Markov chain to recover distribution of interest we need it to be *ergodic* (so that we can apply a more general LLN on the simulated draws). We can spend several pages on the definitions of stochastic processes and get to Birkhoff ergodic theorem, but lets try to use some intuition. Discrete space is more convenient for this, so lets take a discrete process as an example.

In discrete case Markov Chain has a transition probability matrix $\mathbb{P}$: let there be n states of X, then $\mathbb{P}$ is a n by n matrix, where entries i, j is a probability of moving from state i to state j. Thus, the right eigenvector of this matrix is one (as probabilities sum up to one): $\mathbb{P}1 = 1$. Let q be a n-dimentional vector. q is a stationary distribution if it is a left eigenvector associated with a unit eigenvalue:

$$q'\mathbb{P} = q'$$

In order for the Markov chain to recover distribution of interest we need that it visits all states of interest. Assume we are interested in all states in the state space. In this case, we need probability of getting to any state j from any state i being strictly greater than zero (not necessarily in one iteration). Than the smallest invariant event is the state space itself (there is not subset of states getting to which we get stuck), and the chain is ergodic (definition of ergodicity says that under a certain probability measure all invariant events have probability zero or one).

If we know that chain is ergodic, it satisfies

$$\frac{1}{N}\sum_{i=1}^{N} X_i \xrightarrow{a.s} E(X) \tag{3.5}$$

if $E|X| < \infty$

$$E\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - E(X)\right|^2\right) \xrightarrow{m-s} 0 \tag{3.6}$$

if $E|X|^2 < \infty$, where X is a random vector.

In a bit more general formulation, result in (3.5) is known as Pointwise Ergodic Theorem.

3.3 Monte Carlo Markov Chain

Now lets combine Monte Carlo methods with Markov chains. We want to formulate Markov chain on a parameter space. Denote $\pi(\cdot)$ a a stationary distribution of this Markov chain (either discrete, continuous, or a mixture of the two). We want to draw from $\pi(\theta|\theta_i, x)$.

We can start with some θ_0. Then draws $\theta_1|\theta_0 \sim \pi(\theta|\theta_0, x)$. Then $\theta_2|\theta_1 \sim \pi(\theta|\theta_1, x)$. Repeating this N times we get a sample $\theta_0, \ldots, \theta_N$.

Using the result in (3.5), this Markov chain would have a stationary distribution $\pi(\theta)$ and would satisfy the following (for any function $h(\cdot)$, so it also holds for posterior $\pi(\theta|x)$):

$$\lim_{N\to\infty} 1/N \sum_{i=1}^{N} h(\theta_i) = \mathbb{E}^{\pi}(h(\theta))$$

Notice two things about the sequence $\theta_0, \ldots, \theta_N$:

- It depends on the initial conditions (for the first iterations);
- It is not iid.

To account for the first problem researchers sometimes drop first B observations (this is called "burn-in" period). The second point is not a problem (as long as the dependence is not too high) as we can always increase the simulation size, and we know that long-run averages of the draws from the Markov chain would converge to the appropriate integral over the posterior distribution $\pi(\theta|x)$, and that posterior distribution constructed from Markov chain draws would closely approximate the true posterior distribution.

There are several questions that follow:

- What are methods or algorithms for specifying the chains with the right stationary distributions?
- Do this methods produce ergodic chains?
- How long should we run this chains?

The most common methods and algorithms are discussed in the following sections.

3.4 Methods and Algorithms

3.4.1 Gibbs Sampler

Gibbs Sampler takes a step forward from simulating from $\pi(\theta|\theta_i, x)$, but it is somewhat more intuitive. Lets start with an example

Example 3.4.1. *Let*

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

We need to simulate the joint distribution of this.

Straightforward solution would simply to simulate a joint distribution directly as discussed before. In this case iid draws are easily available. Assume, however, that we can simulate only from conditional distributions:

$$\theta_2|\theta_1 \sim N(\rho\theta_1, 1-\rho^2)$$

$$\theta_1|\theta_2 \sim N(\rho\theta_2, 1-\rho^2)$$

Gibbs sampler simulates $\begin{pmatrix} \theta_{1i} \\ \theta_{2i} \end{pmatrix}$ draws in the following manner:

1. Start with some $\begin{pmatrix} \theta_{10} \\ \theta_{20} \end{pmatrix}$;
2. Draw $\begin{pmatrix} \theta_{11} \\ \theta_{21} \end{pmatrix}$ in two steps:

- $\theta_{21} \sim N(\rho\theta_{10}, 1-\rho^2)$
- $\theta_{11} \sim N(\rho\theta_{21}, 1-\rho^2)$

3. Repeat step 2. drawing $\theta_2|\theta_1, \ldots, \theta_N|\theta_{N-1}$ for some N

The resulting draws are highly correlated, but for sufficiently large N produce a good approximation of the joint distribution. The method is quite useless in this case as iid draws are available; however, it is very powerful when iid draws are computationally infeasible.

In more general case, vector of parameters can be partitioned into more than two groups, and the same iterative procedure applies:

1. Start with some $(\theta_{10}, \ldots, \theta_{k0})'$;
2. Draw $(\theta_{11}, \ldots, \theta_{k1})'$ in two steps:
   - $\theta_{11} \sim f_1(\theta_1|\theta_{20}, \ldots, \theta_{k0})$
   - $\theta_{21} \sim f_2(\theta_2|\theta_{11}, \theta_{30}, \ldots, \theta_{k0})$
   - ...
   - $\theta_{k1} \sim f_k(\theta_k|\theta_{11}, \ldots, \theta_{(k-1)1})$
3. Repeat step 2. drawing $\theta_2|\theta_1, \ldots, \theta_N|\theta_{N-1}$ for some N

Notice that f can be the posterior distribution $\pi(\theta_{i+1}|\theta_i, x)$. Hence, Gibbs sampler allows to approximate numerically $\pi(\theta|x)$ by simulating from conditionals (i.e. in case of partitioning θ into two) $\pi(\theta_2|\theta_1, x)$ and $\pi(\theta_1|\theta_2, x)$. This is one way to see that knowing two conditionals implies knowing a joint distribution.

Lets discuss two examples where Gibbs sampler can be more useful. First, we go back to seemingly unrelated regressions; second, we introduce the idea of data augmentation.

3.4.2 Gibbs and SUR

Reconsider the model in (2.10):

$$y = X\beta + \epsilon$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \qquad X = \begin{bmatrix} X_1 & 0 & 0 & 0 \\ 0 & X_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & X_m \end{bmatrix} \qquad \epsilon \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{bmatrix}$$

$$\epsilon \sim N(0, \Sigma \otimes I_n)$$

Lets assume a simple prior structure:

$$\pi(\beta, \Sigma) = \pi(\beta)\pi(\Sigma)$$

$$\beta \sim N(\bar{\beta}, A^{-1})$$

$$\Sigma \sim IW(\nu_0, V_0)$$

This priors are said to be conditionally conjugate: given β we can draw from Σ, and vice versa. Given Σ, lets premultiply the system of equations by the inverse of Cholesky root of Σ, $\Sigma = U'U$, so that error terms would be uncorrelated:

$$\tilde{y} = \tilde{X}\beta + \tilde{\epsilon}$$

$$\tilde{y} = ((U^{-1})' \otimes I_n)y \qquad \tilde{X} = ((U^{-1})' \otimes I_n)X \qquad var(\tilde{\epsilon}) = I_m \otimes I_n$$

Then we can write conditional posteriors

$$\beta|\Sigma, X, y \sim N((\tilde{X}'\tilde{X} + A)^{-1}(\tilde{X}'\tilde{y} + A\bar{\beta}), (\tilde{X}'\tilde{X} + A)^{-1}) \tag{3.7}$$

$$\Sigma|\beta, X, y \sim IW(\nu_0 + n, E'E + V_0) \qquad E = [\epsilon_1, \ldots, \epsilon_m] \tag{3.8}$$

Now implement Gibbs sampler by

- Start with some β_0, Σ_0;
- draw $\beta_1|\Sigma_0$ from (3.7);
- draw $\Sigma_1|\beta_1$ from (3.8);
- Repeat.

3.4.3 Data Augmentation

Think about the following example:

Example 3.4.2. *We are interested in the parameters of a Tobit model. The latent variable is*

$$Y_i^* = X_i'\beta + \epsilon_i$$

where $\epsilon_i|X_i \sim N(0, 1)$. *We observe*

$$Y_i = max(0, Y_i^*)$$

and the regressors X_i. *Suppose prior of* β $\pi(\beta) \sim N(\mu, \Omega)$.

The posterior distribution of β does not have a closed form expression. This is due to the fact that these is no conjugate distribution for this problem. Gibbs sampler comes as a neat solution. Lets treat both $Y^* = (Y_1^*, \ldots, Y_N^*)$ and β as random variables. Conditional distribution of $Y^*|\beta, X$ is

$$Y_i^*|\beta, X \sim TN(X_i'\beta, 1, 0)$$

where $TN(\mu, \sigma^2, c)$ is a truncated normal with c being an upper bound (truncation from above). Conditional distribution of $\beta|Y^*, X$ is

$$\beta|Y^*, X \sim N((X'X + \Omega^{-1})^{-1}(X'Y + \Omega^{-1}\mu), (X'X + \Omega^{-1})^{-1})$$

The latter is a β bayesian estimate in the usual linear regression model. Notice that if $\Omega = \infty$ β corresponds to the asymptotic OLS estimator.

We can apply Gibbs sampler, drawing $Y^*|\beta, X$ and $\beta|Y^*, X$, replacing β and Y^*. Repeating this steps would give us posterior distribution $\pi(\beta|X)$.

The stage of constructing latent variables Y^* is called *data augmenting*.

3.4.4 Hierarchical Models

One of the most common applications of Gibbs sampler is to hierarchical models. Hierarchical models are models constructed from a sequence of conditional distributions. Before, we had a prior $\pi(\theta_1)$ and a likelihood $f(x|\theta_1)$. Lets denote prior as the first step and likelihood as the second step in forming a posterior.

Now, lets add another parameter θ_2 to the prior distribution: $\pi(\theta_1, \theta_2)$. Importantly, likelihood $f(x|\theta_1)$ still depends only on θ_1. We can write this joint distribution as a product of marginal and conditional:

$$\pi(\theta_1, \theta_2) = \pi(\theta_2)\pi(\theta_1|\theta_2)$$

We can think about it as adding another initial step in computing the posterior: now, first step is $\pi(\theta_2)$, second step is $\pi(\theta_1|\theta_2)$, and third step is $f(x|\theta_1)$. First and second steps represent hierarchical structure, and are called first stage and second stage, respectively. Usually θ_2 is of much lower dimensions than θ_1.

I.e., consider simple regression:

$$Y_i = X_i'\beta + \epsilon_i$$

where $\epsilon_i \sim N(0, 1)$. Before we would specify a prior on β, say $\beta \sim N(0, I)$, and compute the posterior given data X and Y. However, specifying a normal distribution for β with fixed mean and variance might be restrictive; instead, we can specify β as $\beta \sim N(0, V_\beta)$, and $V_\beta \sim IW(\nu, V)$, where IW denotes inverted Wishart. This allows for a much more flexible specification of prior distribution of beta.

Hierarchical structure is often used in more complicated settings and estimated with Gibbs sampler. However, for one of the stages of Gibbs drawing a distribution from posterior can be burdensome (i.e. example when we condition on both data and some other parameter). To address this issue we discuss another frequently used algorithm, Metropolis-Hastings. This is a powerful method which can also be coupled with Gibbs sampler (often referred to as hybrid MCMC methods).

3.4.5 Mixture of Normals

A frequently used example of hierarchical structure is mixture of normals. Assume the following prior distribution for y_i:

$$y_i \sim N(\mu_{ind_I}, \Sigma_{ind_i})$$

$$ind_i \sim \text{Multinomial}(pvec)$$

y_i has N dimensions, and $pvec$ is a vector of K probabilities.Hence, for each i y_i can be drawn from one of K Normal distributions, depending on realization of ind_i. This model allow for a very flexible specification; there is a result showing that mixture of normals can approximate any non-parametric dsitribution, adding enough mixture components.

Priors of mixture of normals can be taken in a convenient conditional conjugate form:

$$pvec \sim \text{Dirichlet}(\alpha)$$

$$\mu_k \sim N(\bar{\mu}, \Sigma_k \otimes \alpha_\mu^{-1})$$
$$\Sigma_k \sim IW(\nu, V)$$

This is a hierarchical structure: *pvec* is the first stage, Σ_k and μk is the second stage.

3.4.6 Metropolis-Hastings Algorithm

Gibbs sampler is an enormously useful procedure, but it requires a lot of simulation from conditionals. Sometimes it can be restrictive (for the same reasons as simulating from the original posterior can be restrictive). Consider the case when $\pi(\theta|x)$ is easy to evaluate, but difficult to draw from. For this cases there is a Metropolis class of algorithms, in particular, Metropolis-Hastings.

Suppose we have current value θ_k. As said above it is difficult to draw from $\pi(\theta|\theta_k, x)$ and use standard MCMC methods. The idea is to find a *candidate distribution* $q(\theta|\theta_k, x)$ (which might not depend on θ_k) which (ideally) would be close to $p(\theta|x)$, but from which it would be easy to draw. Then we can draw from this distribution and accept or reject this draws with probability

$$p(\theta, \theta_k) = min(1, \frac{\pi(\theta|x)q(\theta_k|\theta, x)}{\pi(\theta_k|x)q(\theta|\theta_k, x)})$$

so that $P(\theta_{k+1} = \theta) = p(\theta, \theta_k)$.

3.4.7 Metropolis Chain with Importance Sampling

Importance sampling relies on having a good approximation of $\pi(\cdot)$. Usually, this is an asymptotic approximation of the posterior with fattened tails. The same idea can be applied to the Metropolis-Hastings algorithm: candidate function $q(\theta|\theta_k, x)$ is taken using the same ideas, so it is independent of current value θ_k: $q(\theta|\theta_k, x) = q_{imp}(\theta|x)$. In this case acceptation probability becomes

$$p(\theta, \theta_k) = min(1, \frac{\pi(\theta|x)q_{imp}(\theta_k|x)}{\pi(\theta_k|x)q_{imp}(\theta|x)})$$

If q_{imp} is a good approximation of π acceptation probability would be close to one. This implies that the chain would have almost no autocorrelation.

As in the importance sampling, it is necessary that candidate distribution q has fatted tails than target distribution π (as in importance sampling). If it is not the case once the chain wonder off to the tails rejection rate falls, so the chain repeats itself a lot, building up mass.

3.4.8 Metropolis Chains Random Walk

Another particular distribution for q is defined by using random walk to generate proposal values:

$$\theta = \theta_k + \epsilon$$

where $\epsilon \sim N(0, s^2\Sigma)$. This candidate density is symmetric, $q(\theta|\theta_k, x) = q(\theta_k|\theta, x)$. Acceptation probability then becomes

$$p(\theta, \theta_k) = min(1, \frac{\pi(\theta|x)}{\pi(\theta_k|x)})$$

At first glance, this seems to be a very neat solution: we do not require much knowledge about $\pi(\cdot)$ as in case of importance function Metropolis, and the chain can easily navigate in the parametric space. Unfortunately, it is only at first glance: RW Metropolis should be tunes by choosing Σ matrix and appropriate s^2 scaling factor. This requires some prior knowledge about π. Luckily, there are methods to determine the scaling of RW Metropolis (see Rossi et al. 2005, Section 3.10.3).

3.4.9 Hybrid MCMC Methods

Gibbs sampler can be combined with Metropolis algorithm. In the models with hierarchical structure (which we will discuss below) we can easily sample from posterior independent of data, but no so easily sample from posterior depending on data. One of possible solutions is to replace the 'Gibbs' draw in the second case with a Metropolis step. For more details please see Rossi et al. (2005), section 3.12.

3.5 Applications

3.5.1 Demand Estimation: Heterogeneity

Please see the paper of Dube, Hitsch and Rossi (2012) [1]. Authors use mixture of normals to control for heterogeneity of agents. We will cover this during the lecture if we have time left.

Bibliography

[1] Dube, Jean-Pierre, Gunter Hitsch and Peter Rossi (2012), State dependence and alternative explanations for consumer inertia, *RAND Journal of Economics* Vol. 41, No. 3, Autumn 2010 pp. 417445

[2] Hansen, L.P. and T.J. Sargent (2013) *Risk, Uncertainty, and Value*, Lecture notes: U. of Chicago

[3] Imbens G., J.M. Wooldridge (2007) Lecture Notes for NBER Summer School: `http://nber.org/WNE`

[4] Robert, C.P. (2007) *The Bayesian Choice*, Springer

[5] Rossi, P.E., G.M. Allenby and R. McCulloch (2005) *Bayesian Statistics and Marketing*, John Wiley & Sons, Ltd.