12-2016

Monte Carlo Methods in Bayesian Inference: Theory, Methods and Applications

Huarui Zhang
University of Arkansas, Fayetteville

Monte Carlo Methods in Bayesian Inference:
Theory, Methods and Applications

A thesis is submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analytics

by

Huarui Zhang
Shanghai Finance University
Bachelor of Science in Mathematics and Applied Mathematics, 2012

December 2016
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

_____
Dr. Avishek Chakraborty
Thesis Advisor

_____
Dr. Giovanni Petris
Committee Member

_____
Dr. Edward A. Pohl
Committee Member

Abstract

Monte Carlo methods are becoming more and more popular in statistics due to the fast development of efficient computing technologies. One of the major beneficiaries of this advent is the field of Bayesian inference. The aim of this thesis is two-fold: (i) to explain the theory justifying the validity of the simulation-based schemes in a Bayesian setting (why they should work) and (ii) to apply them in several different types of data analysis that a statistician has to routinely encounter. In Chapter 1, I introduce key concepts in Bayesian statistics. Then we discuss Monte Carlo Simulation methods in detail. Our particular focus in on, Markov Chain Monte Carlo, one of the most important tools in Bayesian inference. We discussed three different variants of this including Metropolis-Hastings Algorithm, Gibbs Sampling and slice sampler. Each of these techniques is theoretically justified and I also discussed the potential questions one needs too resolve to implement them in real-world settings. In Chapter 2, we present Monte Carlo techniques for the commonly used Gaussian models including univariate, multivariate and mixture models. In Chapter 3, I focused on several variants of regression including linear and generalized linear models involving continuous, categorical and count responses. For all these cases, the required posterior distributions are rigorously derived. I complement the methodological description with analysis of multiple real datasets and provide tables and diagrams to summarize the inference. In the last Chapter, a few additional key aspects of Bayesian modeling are mentioned. In conclusion, this thesis emphasizes on the Monte Carlo Simulation application in Bayesian Statistics. It also shows that the Bayesian Statistics, which treats all unknown parameters as random variables with their distributions, becomes efficient, useful and easy to implement through Monte Carlo simulations in lieu of the difficult numerical/theoretical calculations.

Acknowledgement

Special thanks to my thesis advisor Dr. Avishek Chakraborty for all of his help with my theses and three courses I have taken. It would be impossible to make it through the semester without his help.

Also, special thanks to my academic advisor Dr. Ed. Pohl for all his help and advice on my study. Without his help, it will be hard for me to finish the whole study smoothly.

Next, special thanks to my first academic year advisor Dr. Giovanni Petris for all his help and advice on my study and two courses I have taken. And without his help, it will be very hard for me to get a good opportunity to implement my knowledge.

Finally, special thanks to my boss Dr. Ed. Gbur for all his help on my two years' study. Without his help, it will be very difficult for me to get through the whole study here.

Then, special thanks are extended to the staff of the University of Arkansas Graduate School for all of their help with thesis and dissertations.

Finally, a special thanks goes out to the faculty and staff at the University of Arkansas for their commitment to the University and to the students.

Table of Contents

List of Figures

List of Tables

Chapter 1: Theory of Bayesian Inference

1.1 Introduction

We begin with a motivation for Bayesian inference for parameter learning from real datasets. Let $D = \{ x_1, x_2, \dots, x_n\}$ be a dataset consisting of n independent and identically distributed (i.i.d.) observations from a distribution $f(x|\theta)$ where θ can be a scalar or vector of unknown parameters. The goal of statistical inference is to learn about θ. Some of the commonly used techniques for this purpose are:

(i) Method of Moments (MoM): equate the empirical moments computed from the sample to the theoretical moments obtained analytically from the definition of f. Then, solve for θ.

(ii) Maximum likelihood (ML): Define the likelihood of the observed sample as a function of unknown θ and solve for θ that maximizes this function:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) \ , \quad \hat{\theta} = \underset{\theta}{argmax}\, L(\theta)$$

$\hat{\theta}$ is referred to as the maximum likelihood estimate (MLE) for θ.

For the above two approaches, one commonality is that we assume θ has one fixed value in reality that we do not know and try to come up an estimate for that single number. In Bayesian inference, we assume θ is also uncertain so it has a probability distribution. The role of the data is to provide information about the probability distribution of θ. Hence, unlike ML and MoM where we try learn about the "true" value of the parameter, in Bayesian inference, we try to learn about its probability distribution.

1.1.1 Prior Distribution

Prior distribution for θ, denoted by $\pi(\theta)$, is reflects our idea about the uncertainty in θ *before* we observe any data. Usually this distribution is constructed using information from previous studies of similar kind and expert scientific knowledge. For example, if we are interested in a model for spending of consumers during the Thanksgiving week of 2016, the consumer spending data from thanksgiving week of past five years can be used to construct the relevant prior distribution. Depending on our idea about the reliability of prior distribution, we can use a highly informative (low-variance) or diffused (large variance) or completely non-informative (all values are equally likely) prior specifications. Most of the time, if we do not have much idea about likely values of a model parameter, we generally use prior distributions with large uncertainty (variance).

We give a very simple example. Suppose $x_1, x_2, \dots, x_n$ are n i.i.d. observations from $N(\theta, 1)$ and our goal is to learn about θ. We may choose $\pi(\theta) = N(\mu_0, \tau^2)$ for some constants μ_0 and τ^2. If we choose τ^2 very small, it reflects our strong belief that the θ is highly likely to be close to μ_0. On the other hand, if we choose $\tau^2 = 10000$, it would mean values of θ far away from μ_0 are also likely. An extreme case would be to choose $\tau^2 = +\infty$, which would mean all real values are equally likely for θ (essentially, $\pi(\theta) = 1(\theta \in \mathbb{R})$), so there is no prior center. Similarly, for a binary dataset $x_1, x_2, \dots, x_n \sim Ber(p)$ assigning $\pi(p) = unif(0,1)$ amounts to a non-informative prior. In some cases, as with the normal example, non-informative priors have an infinite integral, so they are also referred to as improper priors.

When we have more than one parameter (θ is a vector), the prior distribution can be specified as (i) independently on each component on θ or (ii) jointly on the entire vector or (iii) decomposing the joint distribution as product of conditionals as marginal as

$$\pi(\theta_1, \theta_2, \dots, \theta_p) = \pi(\theta_1)\pi(\theta_2|\theta_1) \dots \pi(\theta_p|\theta_1, \theta_2, \dots \theta_{p-1})$$

so that, we can start with a marginal prior for θ_1, conditional prior for θ_2 given θ_1, conditional prior for θ_3 given (θ_1, θ_2) and so on. In Chapters 2 and 3, we show examples of all three kinds of prior specifications.

1.1.2 Hierarchical Model

In a Bayesian inference, whenever we are uncertain about exact values of a parameter, we assign a probability distribution to it. For example, in the above setting of data from $N(\theta, 1)$, one may also treat the two parameters of the prior distribution μ_0 and τ^2 as unknown. In that case, we need to assign a joint probability distribution to these two quantities – we refer to that distribution as hyper-prior (prior on prior parameters). Thus, these parameters and probability distributions can be stacked in a hierarchy with the observed data being at the lowest level.

1.1.3 Posterior Distribution

The posterior distribution reflects the uncertainty in θ *after* we observe the dataset D. The probability model for the data depends on θ through the likelihood function $L(\theta)$. The posterior distribution for θ, denoted by $\pi(\theta|D)$, is the conditional distribution of θ given data calculated as follows using Bayes theorem:

$$\pi(\theta|D) = \frac{L(D|\theta)\,\pi(\theta)}{\int L(D|\theta)\,\pi(\theta)d\theta}$$

It is useful to note that the denominator of above expression is a normalizing constant (free of θ, only a function of data), so we can write:

$$\pi(\theta|D) \propto L(D|\theta)\,\pi(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} * \text{Prior}$$

Posterior distribution involves the observed dataset as conditioning variable. Since the posterior distribution is the centerpiece of Bayesian inference, it must be a proper density. It is useful to remember that use of improper prior distribution is acceptable as long as it does not lead to an improper joint or marginal posterior for one or more parameters.

Most of the time, it is of interest to find a prior having the same functional form as likelihood (when viewed as a function of the parameter), so that the posterior and prior belong to the same family of distributions with different parameters. We refer to such priors as conjugate prior. As we will see in Chapter 2, for any Gaussian dataset, a normal prior for population mean and an Inverse-Gamma prior for population variance will act as conjugate priors.

1.1.4 Posterior Inference

Once we obtain the posterior distribution of θ as above, we can study its properties like posterior mean/median/variance/quantiles by analyzing the function $\pi(\theta|D)$. In general, we can compute $\pi(\theta \in A|D)$ for any region A in the range of θ either analytically or numerically.

For example, if we want to obtain the mean or median of the parameter, we could use:

$$\text{mean} = \int \theta\pi(\theta|D)d\theta \ ; \text{median} = M \text{ where } \int_{\theta \le M} \pi(\theta|D)d\theta = \frac{1}{2}$$

Usually, mean or median is used as point-level summary of the posterior distribution. We may also report an interval within which the parameter lies with a specified probability calculated from its posterior distribution. This is referred to as credible set. For example, an (1-α) credible set for a parameter θ can be defined as the interval $\{\theta: a \le \theta \le b\}$ where

$$\int_a^b \pi(\theta|D)d\theta \ = 1 - \alpha$$

Typically, for real valued parameters, a and b are chosen as $\frac{\alpha}{2}$ and $\left(1 - \frac{\alpha}{2}\right)$ quantiles of the posterior for θ. For non-negative valued parameters (such as scale or precision parameters), one can use credible sets of the form $\{\theta: \theta \le b\}$.

In many situations (as we are going to see later), especially when θ is a vector of parameters (like the location and scale parameters of a normally distributed data where the range of both components of θ are unbounded), it is often difficult to extract the properties of θ as above because of difficulties in solving the problem analytically or numerically. One alternative approach is to follow Monte-Carlo (MC) methods described below.

1.2 Monte Carlo Methods

Monte Carlo methods refer to simulation-based approximation to evaluate analytically intractable integrals of the forms described above. The foundation for Monte Carlo method comes from the law of large numbers that says:

Theorem 1: If X is a random variable with $E|X| < \infty$ and $x_1, x_2, \dots, x_m$ are i.i.d draws from the distribution of X, then as $m \to \infty$

$$\frac{1}{m}\sum_{i=1}^{m} x_i \;\to\; E(X) \quad \text{with probability 1}$$

Hence, if we are asked to compute any integral of the form $\int g(\theta)\pi(\theta|D)d\theta$ for a known integrable function g, we can alternatively simulate a large number of (i.i.d.) observations $\theta_1, \theta_2, \dots, \theta_m$ from posterior density of θ, evaluate the function g at those m points and approximate this integral with $\frac{1}{m}\sum_{i=1}^{m} g(\theta_i)$. We give some examples below.

(a) If our interest is to know the posterior probability of a set $\{\theta: a \le \theta \le b\}$, we rewrite that as

$$\int_a^b \pi(\theta|D)d\theta \;= \int 1(a \le \theta \le b)\pi(\theta|D)d\theta.$$

So, using $g(\theta) = 1(a \le \theta \le b)$ we can repeat the above mentioned steps to approximate this probability.

(b) Consider a vector of parameters: $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ and we want to get posterior correlation between θ_1 and θ_3. We can express this using the following formula:

$$\text{corr}(\theta_1\theta_3) = \frac{\text{cov}(\theta_1\theta_3)}{sd(\theta_1)sd(\theta_3)} = \frac{E(\theta_1\theta_3) - E(\theta_1)E(\theta_3)}{\sqrt{E\left(\theta_1{}^2\right) - E(\theta_1)^2}\sqrt{E\left(\theta_3{}^2\right) - E(\theta_3)^2}}$$

All of the integrals in above expression can similarly be evaluated by using large number of draws from $\pi(\theta_1, \theta_2, \theta_3|D)$ and it turns out to be the correlation coefficient r between posterior samples of θ_1 and θ_3.

The utility of Monte Carlo methods depends on our ability to generate a large number of observations from the target posterior of θ. Next, we discuss some variations of Monte Carlo methods.

1.2.1 Exact Monte Carlo

When θ is a scalar, in many situations, $\pi(\theta|D)$ is standard is easy to directly sample from. A trivial example could be the posterior density of μ when we have n i.i.d observations from $N(\mu, 1)$ and the prior $\pi(\mu) = N(\mu_0, \sigma_0^2)$ is completely specified. It is easy to see that the posterior of μ will also be normal and can be sampled a large number of times efficiently.

When $\pi(\theta|D)$ is not a standard density, one can use techniques like inverse cumulative distribution function (CDF), acceptance-rejection sampling or importance sampling to draw samples from it. While availability of closed form inverse CDF is essential for the first one, the efficiency of the latter two techniques will depend on finding an easy-to-draw-from density function that is reasonably similar to the target density.

Moving to a multi-dimensional θ, drawing from its joint posterior is still possible in some cases (such as mean vector of a multivariate normal with completely known dispersion matrix). Alternatively, one can also draw from the joint posterior by using conditional and marginal draws in succession. To understand this, note that, we can always write:

$$\pi(\theta_1, \theta_2, \ldots \theta_p|D) = \pi(\theta_1|D) \prod_{i=1}^{n} \pi(\theta_i|\theta_{1:(i-1)}, D)$$

Hence, starting with a draw from marginal posterior of θ_1, we draw every other θ_i from its conditional posterior given the previously drawn components of θ. Combining these p univariate draws produces one draw exactly from the joint posterior of θ. We shall see illustration of this technique in Chapters 2 and 3.

In many examples, specifically in those involving complex hierarchical structures, analytically integrating out the components of θ to find the above chain of marginal and conditional distributions can be a difficult task. In those cases, we use a Markov chain to traverse the range of θ. However, using a Markov chain would necessitate one more level of convergence to hold (more on that later). Hence, whenever exact sampling is an option, we must always adopt that.

1.2.2 Markov Chain Monte Carlo

We begin with a few relevant definitions. A sequence of a random variable $\{x^{(0)}, x^{(1)}, \dots\}$ is a Markov Chain if the conditional distribution of $x^{(n)}$, given $x^{(0)}, x^{(1)}, \dots, x^{(n-1)}$, only depends on $x^{(n-1)}$.

If $x^{(0)} \sim f_0(x^{(0)})$, then

$$\mathrm{f}_1\left(x^{(1)}\right) = \int q\left(x^{(1)}|x^{(0)}\right) f_0(x^{(0)})\, dx^{(0)}$$

Here, $q\left(x^{(1)}|x^{(0)}\right)$ is called transition kernel of the Markov chain. In general,

$$f_t\left(x^{(t)}\right) = \int q\left(x^{(t)}|x^{(t-1)}\right) f_{t-1}(x^{(t-1)})\, dx^{(t-1)}$$

If $p_s(x)$ is a probability density function such that $x^{(t)} \sim p_s \Rightarrow x^{(t+1)} \sim p_s$

then $p_s(x)$ is called stationary distribution for the Markov Chain. Obviously, the form of p_s (if it exists) depends on the form of q. If we simulate $x^{(0)} \sim p_s$, all subsequent steps will produce (correlated) samples from p.

Think of a Markov chain on θ with stationary distribution $\pi(\theta|D)$. If we can start with a $\theta^{(0)}$ drawn from this distribution, all subsequent samples will also be draws from the posterior, so all

Monte Carlo based computations can be performed using them. So, we need to address two issues:

(A) How can we find a transition kernel q with stationary distribution same as $\pi(\theta|D)$?

(B) If (A) is resolved, how do we circumvent the requirement that $\theta^{(0)}$ must be drawn from the posterior?

We provide answer to (A) in the next section and focus on (B) here. We present a few concepts and results related to Markov chains for that. See Isaacson and Madsen (1976) for details.

We call a distribution $p_L(x)$ to be the limiting distribution of a Markov chain if,

$$p_L(A) = \lim_{t\to\infty} P(X^{(t)} \in A | X^{(0)} = x^{(0)})$$

does not depend on the initial state $x^{(0)}$. Limiting distribution may or may not exists for a Markov chain.

We call a Markov chain irreducible, if there is a path to go from every state to every other state. We call a Markov chain aperiodic if for any two states a and b, the gcd of all path lengths that go from a to b is 1. We call a Markov chain positive recurrent, if starting from any state, the expected time to return to that state is finite. Now, we state the main result that addresses question B.

Theorem 2: For an ergodic (irreducible, aperiodic and positive recurrent) Markov chain, there exists a limiting distribution which is also its unique stationary distribution.

It implies if we can (answer Question (A) and) find an ergodic Markov chain with stationary distribution $\pi(\theta|D)$, the marginal distribution of draws from that chain will converge to a limiting distribution which is same as the stationary distribution $\pi(\theta|D)$, irrespective of the

distribution of the initial parameter vector $\theta^{(0)}$. This sampling technique is referred to as Markov chain Monte Carlo (MCMC; Gilks 1995).

Suppose $q(\theta^{(t+1)}|\theta^{(t)})$ be a transition kernel with stationary distribution $\pi(\theta|D)$. If we draw $\theta^{(0)}$ from any distribution of our choice (or alternatively set it equal to a fixed value) and keep drawing $\theta^{(1)}, \theta^{(2)}$ as:

$$\theta^{(0)} \xrightarrow{q(\theta^{(1)}|\theta^{(0)})} \theta^{(1)} \xrightarrow{q(\theta^{(2)}|\theta^{(1)})} \theta^{(2)} \rightarrow \cdots$$

then, after a large number of draws N are completed, $\theta^{(N+1)}, \theta^{(N+2)}$, ... can be approximated as correlated samples from $\pi(\theta|D)$. Thus, we need to discard a large number of initial draws, referred to as burn-in period in an MCMC.

How do we remove the correlation? For that, we only collect at draws of the Markov chain at a certain interval d such as $\theta^{(N+1)}, \theta^{(N+d+1)}, \theta^{(N+2d+1)}$, ... Larger the value of d is, weaker is the correlation between the successive observations. This procedure of only using observations at a certain interval is called thinning. An MCMC algorithm typically uses both burn-in and thinning so that the leftover samples approximate as much as possible a set of independent draws from $\pi(\theta|D)$.

1.3 Theory and Methods in MCMC

Now, we focus on exploring options for transition kernel that has $\pi(\theta|D)$ as the stationary distribution. In the following, I will describe three kinds of methods in MCMC with necessary theoretical justification.

1.3.1 Metropolis-Hastings Algorithm

Consider a situation where we have the closed form expression for $\pi(\theta|D)$. We do not know how to sample from it but, given a point θ, we can evaluate it up to a normalizing constant. The Metropolis-Hastings (MH) algorithm works in this scenario by proposing successive values of θ from a proposal distribution g that is completely known and easy to draw from (Chib and Greensburg 1995). Given $\theta^{(i)}$, we can draw $\theta^{(\text{propose})}$ from $g(\theta^{(\text{propose})}|\theta^{(i)})$. So, the recent most state of θ serves as a parameter in g. Then, we calculate an acceptance probability p_A given by

$$p_A = p_{\theta^{(i)} \to \theta^{(\text{propose})}} = \frac{\pi\left(\theta^{(\text{propose})}|D\right)}{\pi(\theta^{(i)}|D)} * \frac{g\left(\theta^{(i)}\middle|\theta^{(\text{propose})}\right)}{g\left(\theta^{(\text{propose})}\middle|\theta^{(i)}\right)} \bigwedge 1$$

Finally, we set the next state of θ as:

$$\theta^{(i+1)} = \begin{cases} \theta^{(\text{propose})} & \text{with probability } p_A \\ \theta^{(i)} & \text{with probability } 1 - p_A \end{cases}$$

One key aspect of MH algorithm is to ensure a reasonable rate of acceptance for the proposals. A good proposal distribution will produce a value of p_A close to 1(so we accept what we propose most of the time). If a proposal distribution produces small values of p_A close to 0 most of the time, the Markov chain of θ often gets stuck at current states and covers only a few states in a long time. In applications, it may be difficult to choose a proposal distribution with large acceptance probability most of the times. Two types of choices are frequent in literature:

(i) Random walk proposal: Propose the new state of θ from a distribution centered at its current state and a small proposal variance. If θ is real valued, we can use $g(\theta^{(\text{propose})}|\theta^{(i)}) \sim N(\theta^{(i)}, \tau^2)$. (Notice that, with this choice of proposal, the ratio involving g disappears from p_A

but that is not generally true when the proposal is not normal.) So, every time we are trying to move a certain distance away in either direction from the present state, similar to the principle of a random walk. If τ is small, we propose in close proximity of the current state, hence we expect p_A to be reasonably large and acceptance rate to go up. But, at the same time, because our moves are small, it may take a long time to traverse the entire domain of θ. Moreover, if the target distribution is multimodal with low and high probability regions mixed with each other, a small proposal variance would make it difficult to propose a direct move from one mode to another without passing through the low probability region (that would mean the move is highly likely to be rejected). Hence, as a result, we may keep moving only within a small sub-domain for a long time.

Choosing τ large would probably reduce the extent of above problem by proposing points that are more scattered across the domain of θ but it would more frequently result in low values of p_A and subsequent rejection. To understand this, notice that p_A depends on the ratio of the posterior at current and proposed states. If the proposed state is far away from current state, it can potentially be in a low posterior probability region and that would make that ratio too small.

Hence, we need to have balance these two conflicting objectives to set a value of τ: efficiently covering the entire domain of θ while ensuring we are not rejecting too many moves. In practical experience, a proposal variance that would result in $30\% - 50\%$ acceptance rate, is reasonable. We can set a target acceptance rate in this region and then increase or decrease τ based on observing a too high or too low acceptance rate. We have presented an example showing the effect of τ on acceptance rate in (refer the figure) in Chapter 2. Harrio et al. (1999) discusses how to use an adaptive proposal distribution in this context.

(ii) Independent proposal: Here, we do not use the current state of θ to propose a new state. So, we can write $g\left(\theta^{(\text{propose})}\middle|\theta^{(i)}\right) = g(\theta^{(\text{propose})})$, free of $\theta^{(i)}$. In that case p_A becomes a function of the ratio of posterior and proposal compared at current and proposed states.

$$p_A = \left[\frac{\pi}{g}\left(\theta^{(\text{propose})}\right)\Big/\frac{\pi}{g}\left(\theta^{(i)}\right)\right] \bigwedge 1$$

The advantage is that the proposal is not connected to what θ currently is, so we can propose values more scattered across the domain. For example, one may use the prior for θ as its proposal and then compute p_A to accept/reject that move. One of the requirements for this to work well is that the high probability regions under the g and the π should not be different. Sometimes, one may use a part of the posterior as proposal so they are not too different in shape. If they are too different, we may end up proposing too many values that have very low values of $\frac{\pi}{g}$ and are likely to be rejected. The Random walk proposal avoids this problem by proposing a move centered at an already accepted stare of θ. See Gåsemyr (2003) on how to choose the independent proposal distribution adaptively.

For practical applications involving exponential families, most often it is computationally efficient to calculate $\log p_A$ and then generate an $\exp(1)$ random number to perform the accept-reject step. This follows from the fact that $u \sim \text{Unif}(0,1) \Rightarrow -\log u \sim \exp(1)$.

Next, we theoretically show that, the Markov chain we proposed here has the target posterior as its stationary distribution. We assert that the required conditions for applying Theorem (refer) is already satisfied. The Markov chain is aperiodic because at every transition, there is a nonzero probability of remaining at the current state. It is clearly irreducible as well.

Basically, this is a Markov Chain with transition kernel $\mathrm{f}\left(\theta^{(i+1)}\middle|\theta^{(i)}\right) * p_{\theta^{(i)}\to\theta^{(i+1)}}$, so $\theta^{(i)}$ can be any value. Then the prerequisite is that $\pi(\theta)$ is the stationary distribution of kernel $\mathrm{q}\left(\theta^{(i+1)}\middle|\theta^{(i)}\right) = \mathrm{f}\left(\theta^{(i+1)}\middle|\theta^{(i)}\right) * p_{\theta^{(i)}\to\theta^{(i+1)}}$. I will prove as follows.

$$\mathrm{p}\left(\theta^{(i)} = \mathrm{a}, \theta^{(i+1)} = \mathrm{b}\right) = \mathrm{p}\left(\theta^{(i)} = \mathrm{a}\right) * \mathrm{p}\left(\theta^{(i+1)} = \mathrm{b}|\theta^{(i)} = \mathrm{a}\right) = \pi(a) * \mathrm{f(b|a)} * p_{\mathrm{a}\to\mathrm{b}}$$

$$\mathrm{p}\left(\theta^{(i)} = \mathrm{b}, \theta^{(i+1)} = \mathrm{a}\right) = \mathrm{p}\left(\theta^{(i)} = \mathrm{b}\right) * \mathrm{p}\left(\theta^{(i+1)} = \mathrm{a}|\theta^{(i)} = \mathrm{b}\right) = \pi(b) * \mathrm{f(a|b)} * p_{\mathrm{b}\to\mathrm{a}}$$

$$p_{\mathrm{a}\to\mathrm{b}} = \left(\frac{\pi(\mathrm{b})}{\pi(\mathrm{a})} * \frac{\mathrm{f(a|b)}}{\mathrm{f(b|a)}}\right) \wedge 1; \;\; p_{\mathrm{b}\to\mathrm{a}} = \left(\frac{\pi(\mathrm{a})}{\pi(\mathrm{b})} * \frac{\mathrm{f(b|a)}}{\mathrm{f(a|b)}}\right) \wedge 1$$

Case-1:

$$\frac{\pi(\mathrm{b})}{\pi(\mathrm{a})} * \frac{\mathrm{f(a|b)}}{\mathrm{f(b|a)}} < 1; p_{\mathrm{a}\to\mathrm{b}} = \frac{\pi(\mathrm{b})}{\pi(\mathrm{a})} * \frac{\mathrm{f(a|b)}}{\mathrm{f(b|a)}}; p_{\mathrm{b}\to\mathrm{a}} = 1$$

$$\mathrm{p}\left(\theta^{(i)} = \mathrm{a}, \theta^{(i+1)} = \mathrm{b}\right) = \mathrm{p}\left(\theta^{(i)} = \mathrm{b}, \theta^{(i+1)} = \mathrm{a}\right) = \pi(\mathrm{b}) * \mathrm{f(a|b)}$$

Case-2:

$$\frac{\pi(\mathrm{b})}{\pi(\mathrm{a})} * \frac{\mathrm{f(a|b)}}{\mathrm{f(b|a)}} > 1; p_{\mathrm{a}\to\mathrm{b}} = 1; p_{\mathrm{b}\to\mathrm{a}} = \frac{\pi(\mathrm{a})}{\pi(\mathrm{b})} * \frac{\mathrm{f(b|a)}}{\mathrm{f(a|b)}}$$

$$\mathrm{p}\left(\theta^{(i)} = \mathrm{a}, \theta^{(i+1)} = \mathrm{b}\right) = \mathrm{p}\left(\theta^{(i)} = \mathrm{b}, \theta^{(i+1)} = \mathrm{a}\right) = \pi(a) * \mathrm{f(b|a)}$$

Case-3

$$\frac{\pi(\mathrm{b})}{\pi(\mathrm{a})} * \frac{\mathrm{f(a|b)}}{\mathrm{f(b|a)}} > 1; p_{\mathrm{a}\to\mathrm{b}} = p_{\mathrm{b}\to\mathrm{a}} = 1; \pi(a) * \mathrm{f(b|a)} = \pi(b) * \mathrm{f(a|b)}$$

$$\mathrm{p}\left(\theta^{(i)} = \mathrm{a}, \theta^{(i+1)} = \mathrm{b}\right) = \mathrm{p}\left(\theta^{(i)} = \mathrm{b}, \theta^{(i+1)} = \mathrm{a}\right)$$

So it is always true that

$$\mathrm{p}\left(\theta^{(i)} = \mathrm{a}, \theta^{(i+1)} = \mathrm{b}\right) = \mathrm{p}\left(\theta^{(i)} = \mathrm{b}, \theta^{(i+1)} = \mathrm{a}\right)$$

if we use kernel $\mathrm{q}\left(\theta^{(i+1)}\middle|\theta^{(i)}\right) = \mathrm{f}\left(\theta^{(i+1)}\middle|\theta^{(i)}\right) * p_{\theta^{(i)}\to\theta^{(i+1)}}$.

$$\pi\left(\theta^{(i+1)}=b\right) = \int q\left(\theta^{(i+1)}=b\middle|\theta^{(i)}=a\right)*\pi\left(\theta^{(i)}=a\right)d\theta^{(i)}$$

$$= \int \mathrm{p}\left(\theta^{(i)}=\mathrm{a},\theta^{(i+1)}=\mathrm{b}\right)d\theta^{(i)} = \int \mathrm{p}\left(\theta^{(i)}=\mathrm{b},\theta^{(i+1)}=\mathrm{a}\right)d\theta^{(i+1)}$$

$$= \int q\left(\theta^{(i+1)}=a\middle|\theta^{(i)}=b\right)*\pi\left(\theta^{(i)}=b\right)d\theta^{(i+1)}$$

$$= \pi\left(\theta^{(i)}=b\right)*\int q\left(\theta^{(i+1)}=a\middle|\theta^{(i)}=b\right)d\theta^{(i+1)} = \pi\left(\theta^{(i)}=b\right)$$

$\theta^{(i)}$ and $\theta^{(i+1)}$ are from identical distribution $\pi(\theta)$, so $\pi(\theta)$ is stationary for kernel $\mathrm{q}\left(\theta^{(i+1)}\middle|\theta^{(i)}\right) = \mathrm{f}\left(\theta^{(i+1)}\middle|\theta^{(i)}\right)*p_{\theta^{(i)}\to\theta^{(i+1)}}$.

1.3.2 Gibbs Sampling

If parameters $\boldsymbol{\theta}$ is a vector $\binom{\theta_1}{\theta_2}$, and posterior distribution of $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta}|\text{data}) = \pi(\theta_1,\theta_2|\text{data})$, which is not standard to simulate from, then we can integrate marginal posterior distributions of θ_1 and θ_2 and they are $\pi(\theta_1|\text{data}) = \int \pi(\theta_1,\theta_2|\text{data})d\theta_2$ and $\pi(\theta_2|\text{data}) = \int \pi(\theta_1,\theta_2|\text{data})d\theta_1$.

$$\pi\left(\theta_1^{(1)}\middle|\text{data}\right) = \int_{\theta_2^{(0)}} \pi\left(\theta_1^{(1)},\theta_2^{(0)}\middle|data\right)\mathrm{d}\theta_2^{(0)} = \int_{\theta_2^{(0)}} \pi\left(\theta_1^{(1)}\middle|\theta_2^{(0)},data\right)*\pi\left(\theta_2^{(0)}\middle|data\right)\mathrm{d}\theta_2^{(0)}$$

$$= \int_{\theta_2^{(0)}} \pi\left(\theta_1^{(1)}\middle|\theta_2^{(0)},data\right)*\left[\int_{\theta_1^{(0)}} \pi\left(\theta_1^{(0)},\theta_2^{(0)}\middle|data\right)\mathrm{d}\theta_1^{(0)}\right]\mathrm{d}\theta_2^{(0)}$$

$$= \int_{\theta_2^{(0)}} \pi\left(\theta_1^{(1)}\middle|\theta_2^{(0)},data\right)*\left[\int_{\theta_1^{(0)}} \pi\left(\theta_2^{(0)}\middle|\theta_1^{(0)},data\right)*\pi\left(\theta_1^{(0)}\middle|data\right)\mathrm{d}\theta_1^{(0)}\right]\mathrm{d}\theta_2^{(0)}$$

$$= \int_{\theta_1^{(0)}} \left[\int_{\theta_2^{(0)}} \pi\left(\theta_1^{(1)}\middle|\theta_2^{(0)},data\right)*\pi\left(\theta_2^{(0)}\middle|\theta_1^{(0)},data\right)d\theta_2^{(0)}\right]*\pi\left(\theta_1^{(0)}\middle|data\right)\mathrm{d}\theta_1^{(0)}$$

$$= \int_{\theta_1^{(0)}} q(\theta_1^{(0)} \to \theta_1^{(1)}) * \pi\left(\theta_1^{(0)} \middle| data\right) \mathrm{d}\theta_1^{(0)},$$

where $q\left(\theta_1^{(0)} \to \theta_1^{(1)}\right) = \int_{\theta_2^{(0)}} \pi\left(\theta_1^{(1)} \middle| \theta_2^{(0)}, \mathrm{data}\right) * \pi\left(\theta_2^{(0)} \middle| \theta_1^{(0)}, \mathrm{data}\right) \mathrm{d}\theta_2^{(0)}$.

$$\theta_1^{(0)} \to \theta_2^{(0)} \to \theta_1^{(1)} \to \theta_2^{(1)} \to \theta_1^{(2)} \ldots$$

Generalized to general situations, kernel q is $\{\pi(\theta_i|\theta_{-i}) : i = 1,2, \ldots, p\}$ (use the recent value). Through this kernel, we can get a Markov Chain with our target distribution as stationary distribution. And this Markov Chain simulation method with full conditional kernel is called Gibbs Sampling (Gelfand 2000). Gibbs Sampling is a special type of Metropolis Hasting Sampling with independent M-H kernel. Acceptance probability $\mathrm{p}_{\theta^{(\mathrm{i})} \to \theta^{(\mathrm{i}+1)}} = \frac{\pi(\theta^{(i+1)})}{\pi(\theta^{(i)})} * \frac{\mathrm{q}(\theta^{(i)})}{\mathrm{q}(\theta^{(i+1)})}$ where $\pi\left(\theta^{(i)}\right) = \pi(\theta^{(i)}|all\ others)$ and $\mathrm{q}\left(\theta^{(i)}\right) = \pi(\theta^{(i)}|\mathrm{all\ others})$. And acceptance probability is equal to 1 always. One drawback of Gibbs Sampling is that the samples will be correlated. The more steps in Gibbs Sampling in multiple parameters $\boldsymbol{\theta}$, the more correlated samples are. To reduce the correlation, if there are multiple parameters to be simulated, we should partition $\boldsymbol{\theta}$ into as few blocks as possible so that it is easy to draw from its joint distribution within each block. $x_1, x_2, \ldots, x_n \sim \mathrm{MVN}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we want to draw from $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathrm{data})$. There are total 9 parameters.

$$\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_3, \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_{12}, \sigma_{13}, \sigma_{23})^T$$

Generally, to simplify the simulation and keep low correlation, people tend to separate them into two partitions.

$$\boldsymbol{\theta}_1 = \mathbf{u} = (\mu_1, \mu_2, \mu_3)^T; \boldsymbol{\theta}_2 = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$$

Then simulate from $\pi(\boldsymbol{\mu}|\boldsymbol{\Sigma}, data)$ and $\pi(\boldsymbol{\Sigma}|\boldsymbol{\mu}, data)$. Table 1 shows how the number of partitions affect the simulation accuracy.

Partition Number	Too Few	Too Many
Advantages	Low or No Correlation	Standard to Draw
Disadvantages	Difficult to Draw	High Correlated

Table 1 Choice of number of blocks in Gibbs sampler

1.3.3 Slice Sampling

Slice sampling (Neal 2003) is a useful tool to draw samples from a posterior distribution which is not standard and not easy to draw from. However, the posterior needs to have certain properties to be suitable for slice sampling. Suppose parameter θ with a target distribution $f(\theta)$. We can implement Slice sampling if following conditions are satisfied. $f(\theta)$ can be written as $f(\theta) = g(\theta)h(\theta)$ with $h(\theta)$ always positive. It is not easy to draw samples from $f(\theta)$, and we do know how to draw from truncated version of $g(\theta)$.

We need to introduce a new random variable in this sampling, say u. And u given θ follows uniform distribution, $u \sim unif(0, h(\theta))$. Then u given θ have probability density function as follows.

$$\pi(u|\theta) = \frac{1}{h(\theta)} * 1(u < h(\theta)) \Longrightarrow$$

$$f(u, \theta) = \pi(u|\theta) * f(\theta) = g(\theta)h(\theta)\frac{1}{h(\theta)} * 1\big(u < h(\theta)\big) = g(\theta) * 1\big(u < h(\theta)\big)$$

$$\Longrightarrow f(\theta|u) = g(\theta) * 1\big(\theta \in H(u)\big)$$

where H is the inverse function of h. We know how to draw samples from $\pi(u|\theta)$ and $f(\theta|u)$, then we can use full conditional distribution to draw samples step by step.

1.4 Predictive Distributions

Suppose, we are given a set of i.i.d. observations $y_1, y_2, \dots, y_n \sim f(\theta)$. How can we predict the possible values of a new observation y_{n+1}? For this, note that the conditional distribution of this new observation given the observed data points can be written as

$$y_{n+1} \sim f(y_{n+1}|y_1, y_2, \dots, y_n) = \int f(y_{n+1}, \theta|y_1, y_2, \dots, y_n)\, d\theta$$

$$= \int f(y_{n+1}|\theta, y_1, y_2, \dots, y_n) f(\theta|y_1, y_2, \dots, y_n)\, d\theta$$

$$= \int f(y_{n+1}|\theta) f(\theta|y_1, y_2, \dots, y_n)\, d\theta$$

This is called posterior predictive distribution of y_{n+1}. In other words, we draw θ's from posterior distribution derived from the data and prior, and then draw one value of y_{n+1} using each simulated θ. We can use these samples to summarize different characteristics of y_{n+1}.

Chapter 2: Bayesian Inference for Gaussian Datasets:

2.1 Introduction

Gaussian distribution, which is also called normal distribution, is one of the most important and common distribution in real world. There is one dimensional and one component normal distribution ($X \sim N(\mu, \sigma^2)$). And there are other more complicated types of normal distributions, such as many dimensional normal distribution ($X \sim MVN_2(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{pmatrix})$), many components normal distribution ($X \sim p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2)$), and many dimensional and many components normal distribution ($X \sim p_1 MVN_2(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{21} & \sigma_2^2\end{pmatrix}) + p_2 MVN_2(\begin{pmatrix}\mu_3\\ \mu_4\end{pmatrix}, \begin{pmatrix}\sigma_3^2 & \sigma_{34}\\ \sigma_{43} & \sigma_4^2\end{pmatrix})$).

2.2 Univariate Normal

$y_1, y_2, \dots, y_n$ come from a normal distribution with mean μ and varianceδ^2 ($y_i \sim N(\mu, \sigma^2)$), and We want to estimate these two parameters from given data. We can easily calculate the mean and variance directly through Maximum Likelihood Estimation, but we want to regard both of them as random variables with some distributions and use Bayesian Method to estimate parameters through Monte Carlo Method. Likelihood of the data is $\text{Likelihood} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y_i-\mu)^2}{2\sigma^2})$. It is not hard to find that the conjugate priors can be normal distribution and inverse gamma distribution (which has been mentioned above) for μ and σ^2 respectively. Then prior for σ^2 is $\sigma^2 \sim IGamma(a_0, b_0)$ and have the probability density function $\pi(\sigma^2) = \frac{{b_0}^{a_0}}{\gamma(a_0)} (\frac{1}{\sigma^2})^{a_0+1} \exp(-\frac{b_0}{\sigma^2})$, where a_0 and b_0 are known. I will do Exact sampling, MCMC sampling, and MH sampling. So for μ, I will put forward two priors $\mu \sim N(\mu_0, \tau_0^2)$ and

$\mu \sim N(\mu_0, c_0\sigma^2)$. The first one is independent prior which I will use in the MCMC sampling and MH sampling, while the second one is dependent prior which I will use in the Exact Sampling. The probability density functions for them are as follows.

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\tau_0^2}}\exp\left(-\frac{(\mu-\mu_0)^2}{2\tau_0^2}\right); \pi(\mu) = \frac{1}{\sqrt{2\pi c_0\sigma^2}}\exp(-\frac{(\mu-\mu_0)^2}{2c_0\sigma^2})$$

the truth is that if we are able to use exact sampling in some distributions, we should prefer the Exact Sampling all the time to other methods.

2.2.1 Dependent prior and exact sampling

Before going into normal estimation, I will introduce a conjugate prior for the variance. Inverse gamma distribution is common treated as the prior of the variance of the normal distribution in Bayesian statistics.

If $x_1, x_2, \ldots, x_n \sim$ N $(0,\sigma^2)$, σ^2 unknown, then

$$\mathrm{L}(\sigma^2) \propto \prod_{i=1}^{n}\frac{1}{\sqrt{\sigma^2}} * \exp\left(-\frac{{x_i}^2}{2*\sigma^2}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} * \exp\left(-\frac{\sum {x_i}^2}{2*\sigma^2}\right) \propto \left(\frac{1}{\sigma^2}\right)^{\alpha-1}$$

$$* \exp\left(-\beta * \left(\frac{1}{\sigma^2}\right)\right)$$

As we can see, the likelihood of the reciprocal of the variance has the form of gamma distribution, and due to this characteristics, the σ^2 have a so-called inverse gamma distribution. So if $\pi(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{a+1} * \exp\left(-b * \left(\frac{1}{\sigma^2}\right)\right)$, then we say $\sigma^2 \sim$ IGamma (a, b). Here, a is called the shape and b is called the scale of the inverse gamma.

And if so, $\frac{1}{\sigma^2}$ ~ Gamma (a, b). Here, a is called the shape parameter, but b is called the rate parameter of the gamma distribution. And then if we want to simulate random numbers through Monte Carlo method for σ^2, then we can simulate Gamma (a, b) and take the reciprocal.

For Exact Sampling in this case, because there are two parameters and we only have the joint distribution, first of all we need to derive the marginal distribution for one of them. After we get the marginal distribution and draw a sample from the marginal distribution, we can use conditional distribution for the other parameter to draw sample. Draw large amount of samples from the above procedure and then calculate the properties of these parameters. I will derive the posterior marginal distribution of μ and conditional distribution of σ^2 next.

Joint Posterior:

$$\pi(\mu, \sigma^2|\text{data}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right) * \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2c_0\sigma^2}\right)$$

$$* \left(\frac{1}{\sigma^2}\right)^{a_0+1} \exp\left(-\frac{b_0}{\sigma^2}\right)$$

$$\Longrightarrow \pi(\sigma^2|\mu, \text{data}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{1}{2}+a_0+1} * \exp\left(-\frac{1}{\sigma^2} * \left(\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2} + \frac{(\mu-\mu_0)^2}{2c_0} + b_0\right)\right)$$

$$\Longrightarrow \pi(\sigma^2|\mu, \text{data}) \sim IGamma\left(\frac{n+1}{2} + a_0, \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2} + \frac{(\mu-\mu_0)^2}{2c_0} + b_0\right)$$

$$\Longrightarrow \pi(\mu|\text{data}) = \int \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{1}{2}+a_0+1} * \exp\left(-\frac{1}{\sigma^2} * \left(\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2} + \frac{(\mu-\mu_0)^2}{2c_0} + b_0\right)\right) d\sigma^2$$

$$= \frac{\gamma\left(\frac{n+1}{2} + a_0\right)}{\left(\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2} + \frac{(\mu-\mu_0)^2}{2c_0} + b_0\right)^{\frac{n+1}{2}+a_0}} \propto \left(\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2} + \frac{(\mu-\mu_0)^2}{2c_0} + b_0\right)^{-\left(\frac{n+1}{2}+a_0\right)}$$

$$\propto \left(1 + \frac{(\mu-D)^2}{E}\right)^{-\left(\frac{n+1}{2}+a_0\right)}; \text{D} = \frac{\mu_0 + c_0\sum y_i}{c_0 n + 1}$$

$$E = \frac{2c_0b_0 + c_0\sum(y_i)^2 + \mu_0{}^2}{c_0n + 1} - (\frac{\mu_0 + c_0\sum y_i}{c_0n + 1})^2$$

Then marginal posterior distribution for μ is T distribution with degree of freedom $v = 2a_0 + n$, location D, and scale $\sigma = \sqrt{\frac{E}{v}}$ (random number $= D + \sigma * t_v, follows\ \pi(\mu|\text{data})$). After draw a sample from the posterior marginal for μ, we can draw a sample from $\pi(\sigma^2|\mu, \text{data})$. There two special cases about the distributions I want to mention. $X \sim N\left(0, \frac{1}{\lambda}\right)$; $\lambda \sim Gamma\left(\frac{r}{2}, \frac{r}{2}\right) \Longrightarrow X \sim t_r$ and $X \sim N(\mu, \sigma^2)$; $\sigma^2 \sim IGamma\left(\frac{r}{2}, \frac{r}{2}\right) \Longrightarrow X \sim t_r\ with\ location\ \mu$.

2.2.2 Independent prior and Gibbs Sampling

For MCMC sampling, we do not need to derive any marginal distribution and are able to draw from full conditional distributions from a Markov Chain one by one until getting a very large sample size.

Joint Posterior:

$$\pi(\mu, \sigma^2|\text{data}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right) * \exp\left(-\frac{(\mu - \mu_0)^2}{2\tau_0^2}\right) * (\frac{1}{\sigma^2})^{a_0+1}\exp(-\frac{b_0}{\sigma^2})$$

$$\Longrightarrow \pi(\mu|\sigma^2, \text{data}) \propto \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right) * \exp\left(-\frac{(\mu - \mu_0)^2}{2\tau_0^2}\right)$$

$$\propto \exp(-\frac{(\mu - (\frac{\tau_0^2\sum y_i + \sigma^2\mu_0}{\sigma^2 + \tau_0^2 n}))^2}{2\,\tau_0^2\sigma^2/(\sigma^2 + \tau_0^2 n)})$$

$$\Longrightarrow \pi(\sigma^2|\mu, \text{data}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right) * \left(\frac{1}{\sigma^2}\right)^{a_0+1} \exp\left(-\frac{b_0}{\sigma^2}\right)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a_0+1} * \exp\left(-\frac{1}{\sigma^2} * (\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2} + b_0)\right)$$

From the derivation above we know both posterior full conditional distributions and can draw from them step by step.

$$\pi(\sigma^2|\mu, \text{data}) \sim IGamma\left(\frac{n}{2} + a_0, \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2} + b_0\right)$$

$$\pi(\mu|\sigma^2, \text{data}) \sim N\left(\frac{\tau_0^2 \sum y_i + \sigma^2 \mu_0}{\sigma^2 + \tau_0^2 n}, \tau_0^2\sigma^2/(\sigma^2 + \tau_0^2 n)\right)$$

As we can see from the mean of the posterior normal distribution for μ, it can be rewritten as

$$\frac{\tau_0^2 n}{\sigma^2 + \tau_0^2 n} * \frac{\sum y_i}{n} + \frac{\sigma^2}{\sigma^2 + \tau_0^2 n} * \mu_0 = \frac{\tau_0^2 n}{\sigma^2 + \tau_0^2 n} * \bar{y} + \frac{\sigma^2}{\sigma^2 + \tau_0^2 n} * \mu_0 = \frac{\sum(w_i * mean)}{\sum w_i}$$

and with n going to infinite we have $\frac{\sum(w_i * mean)}{\sum w_i} \xrightarrow{n \to \infty} \bar{y}$.

2.2.3 Independent prior and MH sampling within Gibbs

To apply Metropolis Hasting Sampling within Gibbs Sampling in univariate normal estimation, the exactly same independent priors for σ^2 and μ from the Gibbs sampling in MCMC will be used. But the difference is that instead of simulating σ^2 from inverse gamma distribution which can be derived from the posterior density function, we propose a dependent log normal distribution for σ^2 to use MH Sampling to simulate σ^2 from the most recent σ^2 value, which can properly deal with the positive property of σ^2. The proposal distribution for σ^2 is $\pi(\sigma^2) \sim LN(\mu_1, \tau_1^2)$ with $\mu_1 = \log(\sigma_{old}^2)$ and τ_1^2 to be some suitable constant keeping the acceptance rate between 30% to 40%. In other words, $\log(\sigma_{new}^2) \sim N(\log(\sigma_{old}^2), \tau_1^2)$. The acceptance probability of the proposal distribution is

$$p_{\sigma_{old}^2 \to \sigma_{new}^2} = \left(\frac{\pi(\sigma_{new}^2|\mu, data)}{\pi(\sigma_{old}^2|\mu, data)} * \frac{q(\sigma_{old}^2|\sigma_{new}^2)}{q(\sigma_{new}^2|\sigma_{old}^2)}\right) \wedge 1$$

Similar to Gibbs Sampling in MCMC, the posterior density of σ^2 is

$$\pi(\sigma^2|\mu, \text{data}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a_0+1} * \exp\left(-\frac{1}{\sigma^2} * (\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2} + b_0)\right)$$

The density of proposal distribution are as follows.

$$\pi(\sigma^2_{new}|\sigma^2_{old}) \propto \frac{1}{\sigma^2_{new}} \exp\left(-\frac{[\log(\sigma^2_{new}) - \log(\sigma^2_{old})]^2}{2\tau_1^2}\right)$$

$$\pi(\sigma^2_{old}|\sigma^2_{new}) \propto \frac{1}{\sigma^2_{old}} \exp(-\frac{[\log(\sigma^2_{old}) - \log(\sigma^2_{new})]^2}{2\tau_1^2})$$

$$p_{\sigma^2_{old}\to\sigma^2_{new}} = \left(\frac{\pi(\sigma^2_{new}|\mu,data)}{\pi(\sigma^2_{old}|\mu,data)}\frac{\sigma^2_{new}}{\sigma^2_{old}}\right) \wedge 1 = (\frac{\left(\frac{1}{\sigma^2_{new}}\right)^{\frac{n}{2}+a_0} * \exp\left(-\frac{1}{\sigma^2_{new}} * (\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}+b_0)\right)}{\left(\frac{1}{\sigma^2_{old}}\right)^{\frac{n}{2}+a_0} * \exp\left(-\frac{1}{\sigma^2_{old}} * (\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}+b_0)\right)}) \wedge 1 \text{ with}$$

$\frac{q(\sigma^2_{old}|\sigma^2_{new})}{q(\sigma^2_{new}|\sigma^2_{old})} = \frac{\sigma^2_{new}}{\sigma^2_{old}}$. Then use full conditional posterior distribution for μ to draw μ conditional on σ^2. Use log normal proposal distribution to draw new σ^2 based on the most recently σ^2, and use acceptance probability derived above to decide if reject or accept the new σ^2 compared to a uniform random number within (0, 1). Then continue to do these steps large amount of times to realize a MH Sampling within Gibbs Sampling in MCMC.

2.3 Mixture Normal

There are k normal distributions with different means and variance, $\{N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), \dots, N(\mu_k, \sigma_k^2)\}$. Probability density function of X is

$$\pi(y) = \pi_1 N(y|\mu_1, \sigma_1^2) + \pi_2 N(y|\mu_2, \sigma_2^2) + \cdots + \pi_k N(y|\mu_k, \sigma_k^2) = \sum_{i=1}^{k} \pi_i N(y|\mu_i, \sigma_i^2)$$

with $\sigma_i^2 > 0, and\ \sum_{i=1}^{k} \pi_i = 1$. And any distribution can be modeled as a mixture of infinitely many normal distributions ($N(\mu, \sigma^2)$'s) with different sets of parameters. $\pi(y) = \sum_{i=1}^{+\infty} \pi_i N(y|\mu_i, \sigma_i^2)$ with $\sum_{i=1}^{k} \pi_i = 1$ ($k \to +\infty$). The parameters we need to estimate are $\{\pi_1, \pi_2, \dots, \pi_k\}$, $\{\mu_1, \mu_2, \dots, \mu_k\}$, and $\{\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2\}$. The prior for each component of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$

are normal and inverse gamma distributions respectively. We will use Dirichlet Distribution with parameter $\boldsymbol{\alpha}_0$ as the prior for probabilities $\boldsymbol{\pi}$.

$$\boldsymbol{\pi}\sim\text{Dir}\big(\boldsymbol{\alpha}_0=(\alpha_1,\alpha_2,\dots,\alpha_k)\big);\pi(\mathbf{P})\propto\pi_1^{\alpha_1-1}*\pi_2^{\alpha_2-1}*\dots*\pi_k^{\alpha_k-1}$$

with $\sum_{i=1}^{k}\pi_i=1$. Because it is not possible or extremely hard to draw samples from $\pi(\boldsymbol{\pi}|\boldsymbol{\mu},\boldsymbol{\sigma}^2,data)$, which is not standard, we will use latent variable **Z.** For observation j, define z_j ($z_j\in\{1,2,\dots,k\}$) represents the component the observation comes from the mixture normal. For example, $z_2=3$ means $y_2\sim N(\mu_3,\sigma_3^2)$. And then we have conditional likelihood, given z=i, is $\pi(y|z=i)=N(y|\mu_i,\sigma_i^2)$ with prior for **Z** is $P(Z=i)=\pi_i$. Hence likelihood

$$\pi(y)=\prod_{j=1}^{n}\pi(y_j|z_j=i)P(Z=i)=\prod_{j=1}^{n}\pi_i N(y_j|\mu_i,\sigma_i^2)$$

, from which after integrating out of Z we can get the actual marginal distribution of Y. Using hierarchical method to write out posterior distribution of all parameters layer by layer.

$\pi(y|z=i)=N(y|\mu_i,\sigma_i^2)$, $P(Z=i)=\pi_i$, $\pi(\mu_i|\sigma_i^2)\sim N(\mu_0,c_0\sigma^2)$ (use dependent prior here), $\pi(\sigma_i^2)\sim IG(a_0,b_0)$, and $\boldsymbol{\pi}\sim\text{Dir}(\alpha_1,\alpha_2,\dots,\alpha_k)$. Joint posterior distribution and conditional posterior distributions are derived as follows.

$$\pi(\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\pi},\boldsymbol{\sigma}^2|data)$$

$$\propto\prod_{j=1}^{n}N\left(y_j\middle|\mu_{z_j},\sigma_{z_j}^2\right)*\prod_{j=1}^{n}\pi_{z_j}*\prod_{i=1}^{k}N(\mu_i|\sigma_i^2)$$

$$*\prod_{i=1}^{k}IG(\sigma_i^2)*Dir(\boldsymbol{\alpha}_0)$$

$$\pi(\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi},\boldsymbol{\sigma}^2,data)\propto\prod_{j=1}^{n}N\left(y_j\middle|\mu_{z_j},\sigma_{z_j}^2\right)*\prod_{j=1}^{n}\pi_{z_j};\pi\left(z_j\middle|\boldsymbol{\mu},\boldsymbol{\pi},\boldsymbol{\sigma}^2,data\right)$$

$$\propto N\left(y_j\middle|\mu_{z_j},\sigma_{z_j}^2\right)*\pi_{z_j}$$

To get posterior distribution for z of each observation j, we have:

$$P(z_j = 1) \propto N(y_j|\mu_1, \sigma_1^2)\pi_1;\ P(z_j = 2) \propto N(y_j|\mu_2, \sigma_2^2)\pi_2; \dots$$

$$P(z_j = k) \propto N(y_j|\mu_k, \sigma_k^2)\pi_k$$

$$P(z_j = 1) = \frac{N(y_j|\mu_1, \sigma_1^2)\pi_1}{\sum_{i=1}^{k} N(y_j|\mu_i, \sigma_i^2)\pi_i};$$

$$P(z_j = t) = \frac{N(y_j|\mu_t, \sigma_t^2)\pi_t}{\sum_{i=1}^{k} N(y_j|\mu_i, \sigma_i^2)\pi_i};\ \sum_{t=1}^{k} P(z_j = t) = 1$$

This is a posterior multinomial distribution with $P(z_j = t)$ mentioned above for **Z**. And then I will derive the posterior distribution for $\boldsymbol{\pi}$ as Dirichlet Distribution as follows.

$$\pi(\boldsymbol{\pi}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, data) \propto \left(\prod_{j=1}^{n} \pi_{z_j}\right) * \pi(\pi_1, \pi_2, \dots, \pi_k)$$

$$\propto \left(\prod_{i=1}^{k} \pi_i^{n_i}\right) * {\pi_1}^{\alpha_1 - 1} * {\pi_2}^{\alpha_2 - 1} * \dots * {\pi_k}^{\alpha_k - 1}$$

$$\propto {\pi_1}^{\alpha_1 + n_1 - 1} * {\pi_2}^{\alpha_2 + n_2 - 1} * \dots * {\pi_k}^{\alpha_k + n_k - 1} \sim Dir(\alpha_1 + n_1, \alpha_2 + n_2, \dots, \alpha_k + n_k)$$

Where n_t represents number of observations which fall in category t. Given z values, posterior distributions for each component of the k normal distributions are similar to the normal distribution simulation before, and the only difference is that we will only use the data belonging to the specific categories. We have

$$\pi(\boldsymbol{\mu}|\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2, data) \propto \prod_{j=1}^{n} N\left(y_j\middle|\mu_{z_j}, \sigma_{z_j}^2\right) * \prod_{i=1}^{k} N(\mu_i|\sigma_i^2)$$

and then we have

$$\pi(\mu_i|\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\sigma}^2, data) \propto \sum_{j:z_j = i} N(y_j|\mu_i, \sigma_i^2) * N(\mu_i|\sigma_i^2)$$

Similarly, we can get

$$\pi(\sigma_i^2|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\pi}, data) \propto \sum_{j:z_j = i} N(y_j|\mu_i, \sigma_i^2) * N(\mu_i|\sigma_i^2) * IG(\sigma_i^2)$$

In generally, k is unknown, and we need to determine value k and do the k-means clustering first. Also for k different normal distributions, the set of parameters $\{z_i, \pi_i, \sigma_i^2\}$ for each iteration is not identifiable. Rearrangement of the sets of parameters need to be done according to some consistent methods, such as order according to $\pi_1, \pi_2, \dots, \pi_k$ and order according to $\mu_1, \mu_2, \dots, \mu_k$ in all the iterations.

2.4 Multivariate Normal

Sequence of random vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$ (each of whom is p dimensional random vector) follows multivariate normal distribution. In other words, $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \sim MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where dispersion matrix $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ is positive definite. To get the posterior distribution, we need to propose the priors for parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively. For $\boldsymbol{\mu}$, it is not difficult find that multivariate normal distribution is a conjugate prior, and we have $\boldsymbol{\mu} \sim MVN_p(\boldsymbol{\mu}_0, \lambda_0 \boldsymbol{\Sigma})$. When talking about $\boldsymbol{\Sigma}$, we need to apply a Wishart distribution to be the prior. Density function and parameters of Wishart distribution are shown as follows.

$$\boldsymbol{\Phi}_{p*p} \sim Wishart_p(d, \boldsymbol{A}); \pi(\boldsymbol{\Phi}) \propto |\boldsymbol{\Phi}|^{\frac{d-p-1}{2}} \exp\left(-\frac{1}{2} trace(\boldsymbol{A}^{-1}\boldsymbol{\Phi})\right); d > p-1$$

To make the posterior distribution easy to be obtained, we will use $\boldsymbol{\Phi} = \boldsymbol{\Sigma}^{-1}$ as our parameter. Then we have $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \sim MVN_p(\boldsymbol{\mu}, \boldsymbol{\Phi}^{-1})$ and $\boldsymbol{\mu} \sim MVN_p(\boldsymbol{\mu}_0, \lambda_0 \boldsymbol{\Phi}^{-1})$. And we have $\boldsymbol{\Phi} \sim Wishart_p(d_0, \boldsymbol{A}_0)$. Then I will derive the posterior distribution of both parameter sets following.

$$\text{Likelihood} \propto |\boldsymbol{\Phi}|^{\frac{n}{2}} \exp\{-\frac{1}{2}\sum_{i=1}^{n}\{(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Phi}(\boldsymbol{x}_i - \boldsymbol{\mu})\}\}$$

$$\pi(\boldsymbol{\mu}) \propto |\boldsymbol{\Phi}|^{\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \frac{\boldsymbol{\Phi}}{\lambda_0}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\}$$

$$\sum_{i=1}^{n}\{(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Phi}(\boldsymbol{x}_i-\boldsymbol{\mu})\}=\sum_{i=1}^{n}\{(\boldsymbol{x}_i-\overline{\boldsymbol{x}}+\overline{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Phi}(\boldsymbol{x}_i-\overline{\boldsymbol{x}}+\overline{\boldsymbol{x}}-\boldsymbol{\mu})\}$$

$$=\sum_{i=1}^{n}\{(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\boldsymbol{\Phi}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})\}+n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Phi}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})$$

$$\pi(\boldsymbol{\mu}|\boldsymbol{\Phi},data)\propto\exp\left(-\frac{1}{2}n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^T\boldsymbol{\Phi}(\overline{\boldsymbol{x}}-\boldsymbol{\mu})\right)\exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\frac{\boldsymbol{\Phi}}{\lambda_0}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right)$$

$$\pi(\boldsymbol{\mu}|\boldsymbol{\Phi},data)\sim MVN_p\left(\frac{n\overline{\boldsymbol{x}}+\frac{\boldsymbol{\mu}_0}{\lambda_0}}{n+\frac{1}{\lambda_0}},((n+\frac{1}{\lambda_0})\boldsymbol{\Phi})^{-1}\right)$$

$$(\boldsymbol{\Phi}|\boldsymbol{\mu},\text{data})\propto\exp\left(-\frac{1}{2}trace(\boldsymbol{L}\boldsymbol{\Phi})\right)*|\boldsymbol{\Phi}|^{\frac{n+d+1-p-1}{2}};\boldsymbol{L}=\boldsymbol{S}+\boldsymbol{B}+\boldsymbol{C}+\boldsymbol{A}^{-1}$$

$$\boldsymbol{B}=n(\overline{\boldsymbol{x}}-\boldsymbol{\mu})(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^T;\boldsymbol{S}=\sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})\{(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\};\boldsymbol{C}=\frac{(\boldsymbol{\mu}-\boldsymbol{\mu}_0)(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T}{\lambda_0}$$

$$\pi(\boldsymbol{\Phi}|\boldsymbol{\mu},\text{data})\sim Wishart_p(n+d+1,\boldsymbol{L}^{-1})$$

From the above full conditional posterior distributions, we can draw samples for all the model parameters.

2.5 Data Analysis

We implement the above sampling schemes discussed above with two real datasets and report the posterior summaries using tables and diagrams.

2.5.1 Iris Dataset

This dataset (Anderson 1935) includes measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. For our purpose, we utilize only a part of the dataset corresponding to the third species *Iris*

Virginica. All four variables for this species satisfies assumption of Gaussian distribution according to Shapiro-Wilk test of normality (Shapiro and Wilk 1965).

Assume Sepal Width in the data set follows normal distribution, say Sepal Width$\sim N(\mu, \sigma^2)$. I estimate parameters with MCMC Gibbs sampling, Exact Sampling, and MCMC MH sampling respectively. First of all, in MH sampling I choose 30 different variances (the acceptance rate for 30 variances shown in Figure 1) for proposed kernel, and choose the one giving approximately 36% acceptance rate with variance 0.4096.

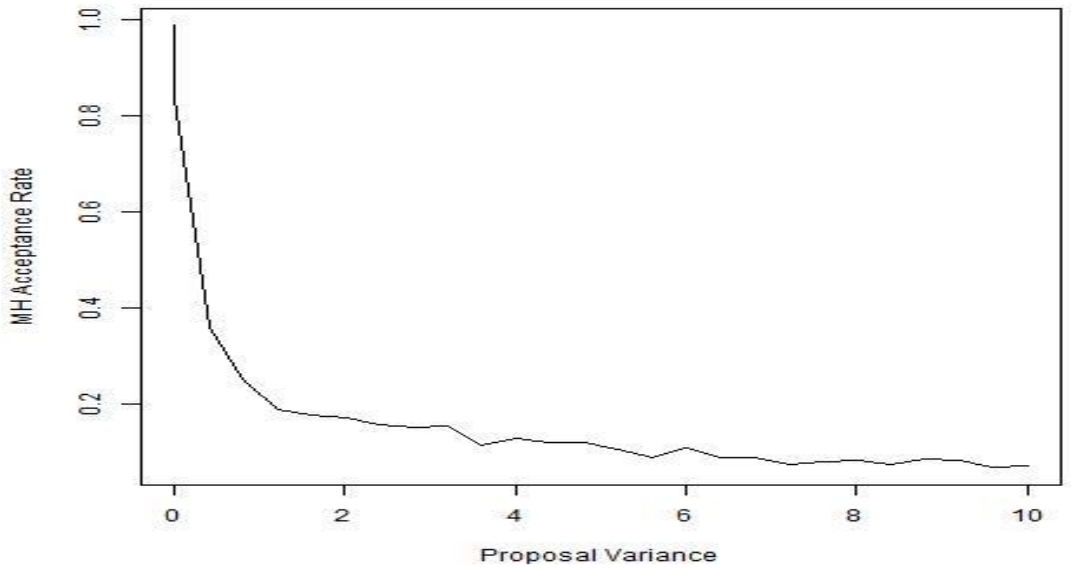

Figure 1. MH variance selection for univariate Normal

Then I simulated the mean μ and variance σ^2 from the normal distribution with three different methods. And the simulation results and summary for parameters posterior distributions are shown in Figure 2 and Table 2 as follows.

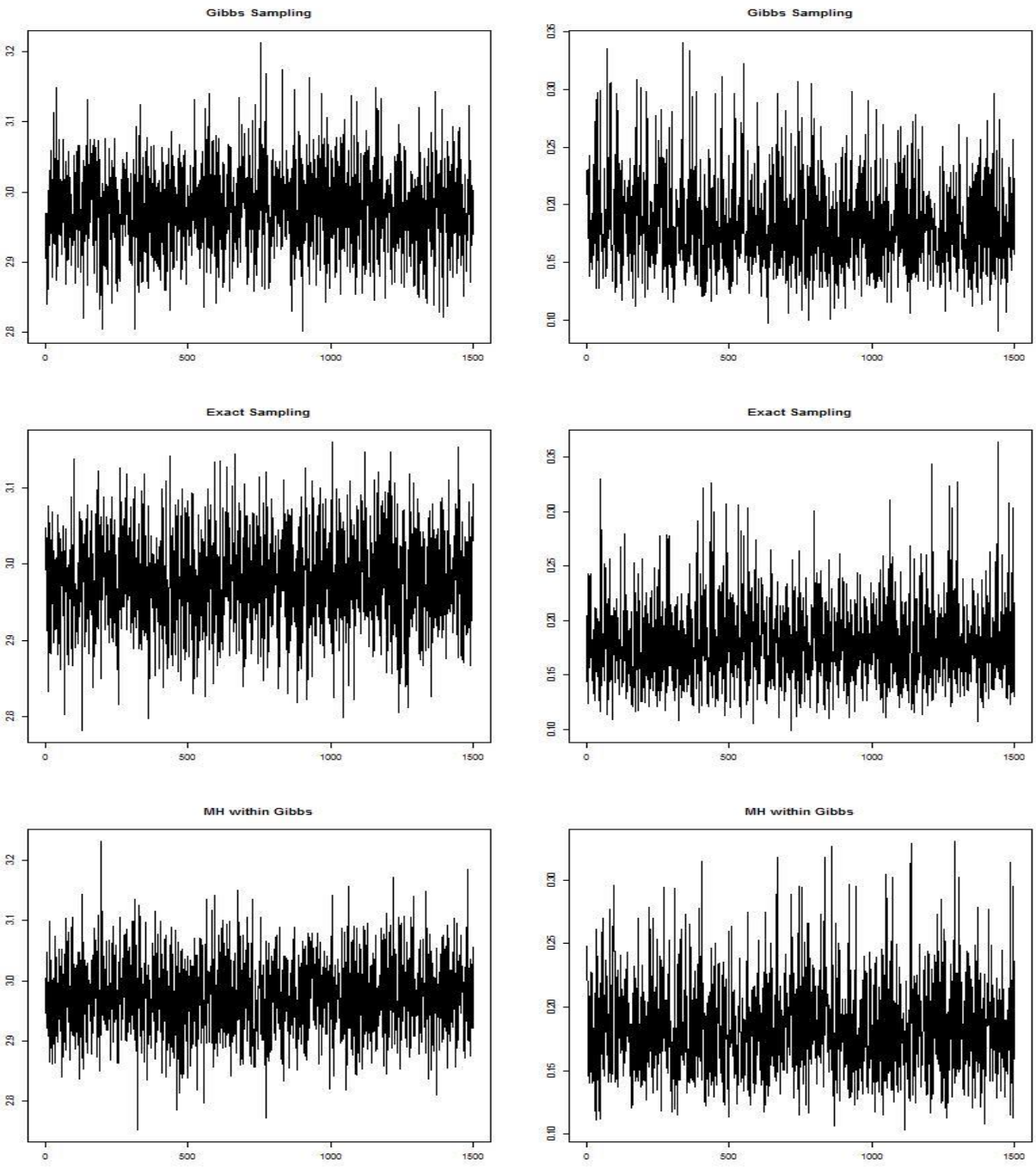

Figure 2. Univariate Normal simulation with three methods

	Mean (μ)	95% credible set μ	Mean (σ^2)	95% credible set σ^2
MCMC	2.974608	(2.861793,3.086442)	0.181001	(0, 0.2502999)
EXACT	2.976459	(2.860402,3.097457)	0.1789585	(0, 0.2431916)
MH	2.974037	(2.857938,3.095774)	0.1832657	(0, 0.2485039)

Table 2 Posterior summary for univariate Normal

Next, I consider (Sepal Length, Sepal Width, Petal Length, Petal Width) as a random vector following multivariate normal distribution, of which each variable follows univariate normal distribution. In other words, $\mathbf{x} \sim MVN_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ and $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ which is a 4 by 4 positive definite symmetric dispersion matrix. And what I will do next is to estimate mean and dispersion matrix for this 4 dimensional multivariate normal distribution with method I have mentioned above. The simulation result and posterior distributions summaries of 4 components of $\boldsymbol{\mu}$ are shown in Figure 3 and Table 3. The simulation result and posterior distribution summaries of components of $\boldsymbol{\Sigma}$ are shown in Figure 4 and Table 4.

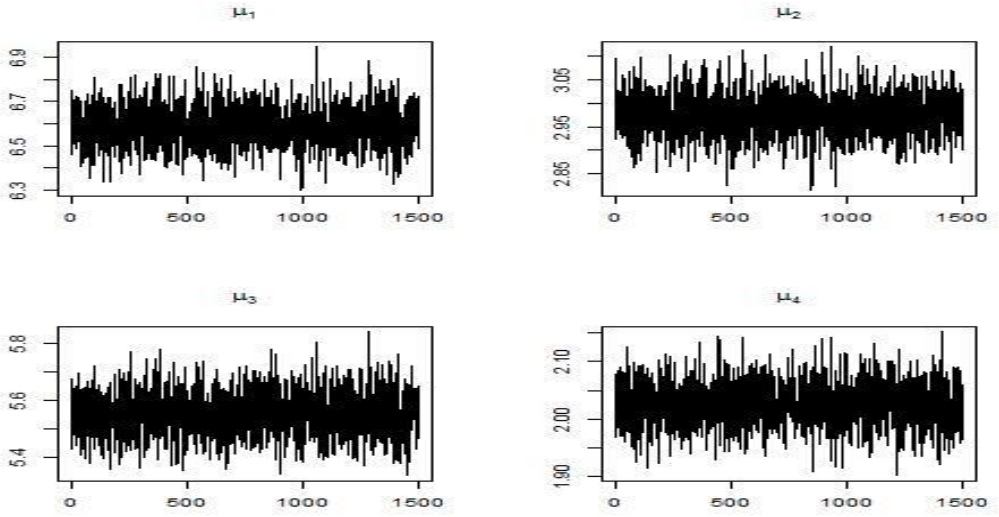

Figure 3. Multivariate Normal mean posterior simulation

	Mean	95% Credible Set
μ_1 (Sepal Length)	6.588038	(6.41531, 6.76037)
μ_2 (Sepal Width)	2.973634	(2.885054, 3.063353)
μ_3 (Petal Length)	5.551997	(5.400633, 5.705720)
μ_4 (Petal Width)	2.026225	(1.947174, 2.099770)

Table 3 Posterior summary for mean in multivariate Normal

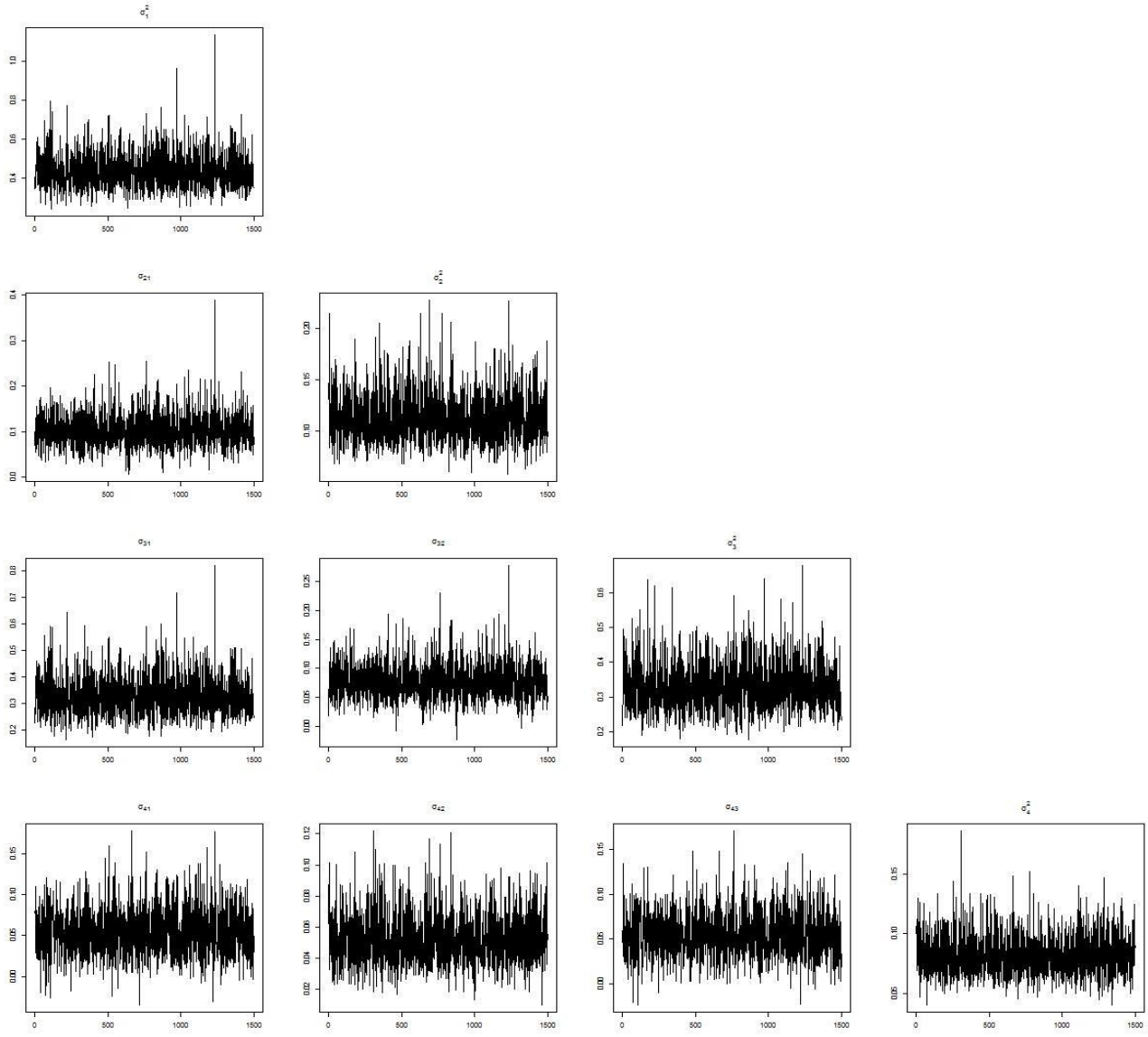

Figure 4. Multivariate Normal dispersion matrix posterior simulation

0.4353 (0.2949, 0.6287)	0.1015 (0.0416, 0.1796)	0.3269 (0.2123, 0.4927)	0.0539 (0.0045, 0.1161)
0.10152 (0.0416, 0.1796)	0.1123 (0.0755, 0.1689)	0.07725 (0.0261, 0.1427)	0.0518 (0.0251, 0.0890)
0.3269 (0.2123, 0.4927)	0.0772 (0.0261, 0.1427)	0.3292 (0.2198, 0.4826)	0.0538 (0.0096, 0.1053)
0.05395 (0.0045, 0.1161)	0.0518 (0.0251, 0.0890)	0.0538 (0.0096, 0.1053)	0.0818 (0.0557, 0.1196)

Table 4 Posterior summary for dispersion matrix in multivariate Normal

In table 4, first line of the cell is mean, and the second line is 95% credible set.

2.5.2 Old Faithful Dataset

This dataset (Azzalini and Bowman, 1990) consists of 272 observations on eruptions from the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. Each observation has two measurements: time duration for an eruption and waiting time for the next eruption, both measured in minutes. We focus only on modeling the density for the latter one. The histogram of the data (refer to Figure) shows a bimodal distribution indicating the potential for using a mixture normal distribution with two components.

Looking at the histogram (refer figure), we decide use two components mixture normal distribution based on some prior knowledge, which is

$$\mathrm{y} = p_1 N(\mu_1, \sigma_1^2) + p_2 N(\mu_2, \sigma_2^2) \quad ; \quad p_1 + p_2 = 1$$

And the simulation results plot and summary of the posterior distributions of parameters are shown in the Figure 5 and Table 5 respectively. Figure 6 shows the probabilities for each observation coming from the first normal component, and it also shows the proportion of times any particular observation was assigned to first component. Figure 7 shows the density of the

data and the estimated density of the mixture normal. From the picture, we can find that the two components mixture normal distribution can well represent the data.

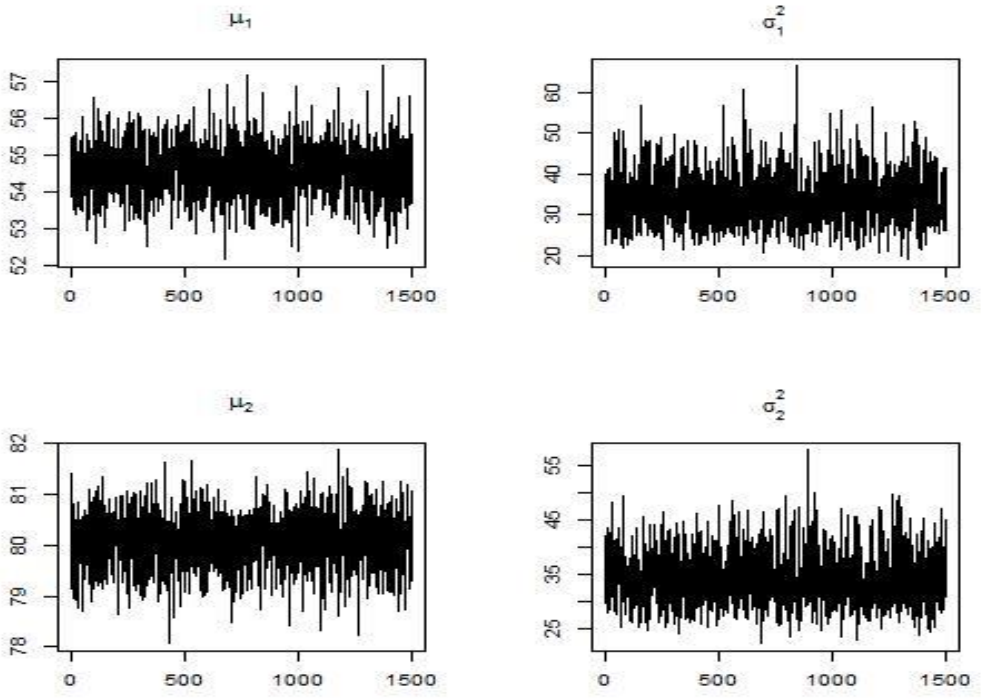

Figure 5. Component-wise mean and variance simulation in mixture Normal

	μ_1	σ_1^2	μ_2	σ_2^2	π_1
Mean	54.57479	33.87583	80.07864	34.51825	0.3636252
95% Credible set	(53.20014, 56.11021)	(0, 45.98382)	(79.06056, 80.98137)	(0, 42.69847)	(0.3060050, 0.4217375)

Table 5 Posterior summary in mixture Normal

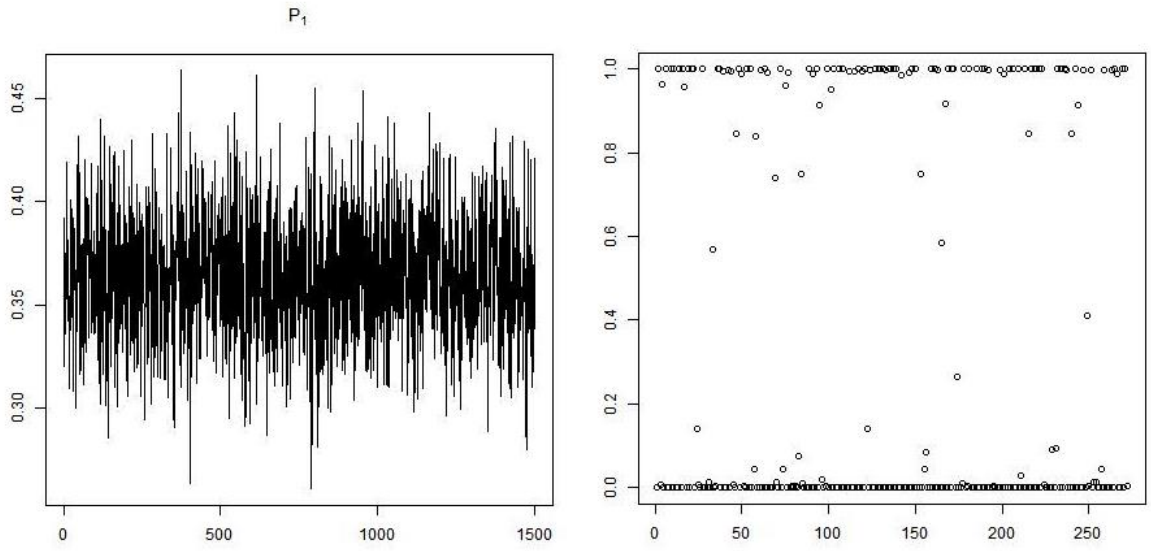

Figure 6. Component probability and indicators simulation in mixture Normal

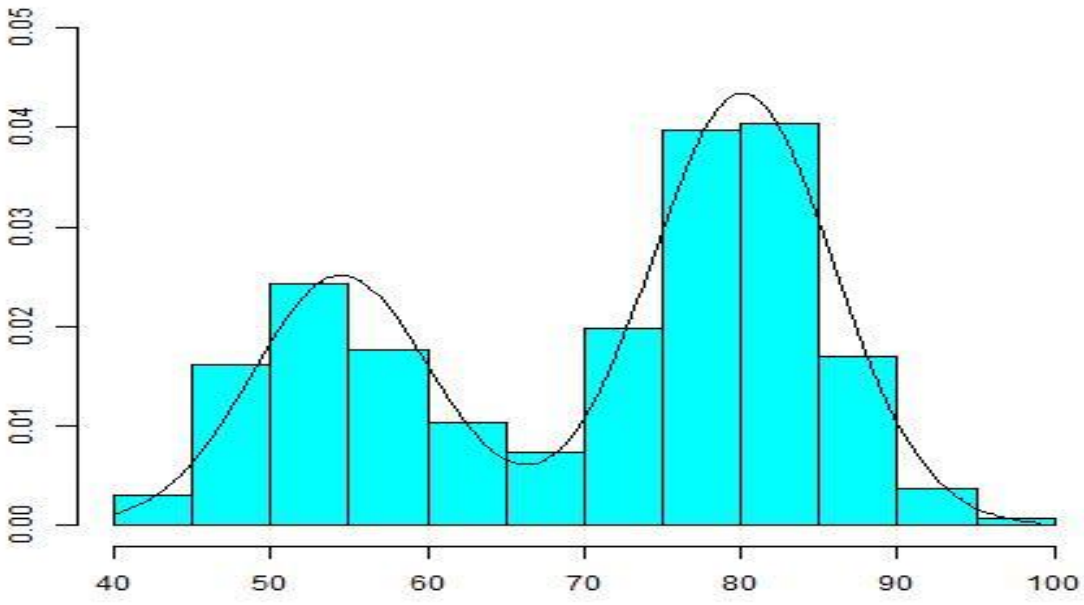

Figure 7. Posterior estimate of mixture Normal density

Chapter 3: Bayesian Inference in Regression

3.1 Introduction

Now, we turn our focus to another common area of data analysis- regression. Regression is a useful tool for many real-world problems whenever we want to find relationships between different variables or try to predict one of them using the value of other variables. In the following, we deal with several different kinds of regression.

3.2 Linear Regression

In normal distribution, $y_1, y_2, \dots, y_n \sim N(\mu, \sigma^2)$, and we can rewrite y_i in another way of regression as $y_i = \mu * 1 + \varepsilon\ ;\ \varepsilon \sim N(0, \sigma^2)$. Adding covariates to the regression and making it general, we have $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i\ ;\ \varepsilon_i \sim N(0, \sigma^2)$, which also be considered as the normal distribution with mean dependent on covariates, as $\mathbf{Y} \sim \mathrm{N}(\mathbf{X\beta}, \sigma^2)$. In this linear regression, the parameters we need to estimate are $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ and σ^2. For σ^2, use the same inverse gamma prior distribution $\sigma^2 \sim IGamma(a_0, b_0)$. But for $\boldsymbol{\beta}$, we will have different forms of multivariate normal distributions, $\boldsymbol{\beta} \sim MVN_p(\boldsymbol{\beta}_0, c_0\sigma^2 \boldsymbol{I}_p)$ and $\boldsymbol{\beta} \sim MVN_p(\boldsymbol{\beta}_0, \tau_0^2 \boldsymbol{I}_p)$ for Exact Sampling and MCMC respectively.

3.2.1 Dependent Prior and exact sampling

To use Exact Sampling in this case I will derive the marginal distribution of $\boldsymbol{\beta}$ and conditional distribution of σ^2. After drawing a sample for $\boldsymbol{\beta}$, draw a sample for σ^2 conditional on $\boldsymbol{\beta}$. Before going into calculation of the marginal and conditional, let's look at the prior and likelihood we have.

$$\text{Likelihood} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{(\mathbf{Y}-\mathbf{X\beta})^T(\mathbf{Y}-\mathbf{X\beta})}{2\sigma^2}\right)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta})}{2\sigma^2}\right)$$

$$Prior\ \pi(\sigma^2) = \frac{{b_0}^{a_0}}{\gamma(a_0)}\left(\frac{1}{\sigma^2}\right)^{a_0+1}\exp\left(-\frac{b_0}{\sigma^2}\right) \propto \left(\frac{1}{\sigma^2}\right)^{a_0+1}\exp\left(-\frac{b_0}{\sigma^2}\right)$$

$$\text{Prior } \pi(\boldsymbol{\beta}) = \left(\frac{1}{2\pi c_0\sigma^2}\right)^{p/2}\exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2c_0\sigma^2}\right) \propto \left(\frac{1}{\sigma^2}\right)^{p/2}\exp\left(-\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2c_0\sigma^2}\right)$$

$$Posterior\ \pi(\boldsymbol{\beta}, \sigma^2|data)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{(n+p)/2+a_0+1} \exp\left(-\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta})}{2\sigma^2} - \frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2c_0\sigma^2} - \frac{b_0}{\sigma^2}\right)$$

$$\Longrightarrow \pi(\sigma^2|\boldsymbol{\beta}, data) \sim IG\left((n+p)/2 + a_0, \frac{(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta})}{2} + \frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2c_0} + b_0\right)$$

$$\Longrightarrow \pi(\boldsymbol{\beta}|data)$$

$$\propto \int \left(\frac{1}{\sigma^2}\right)^{(n+p)/2+a_0+1}$$

$$* \exp\left(-\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta})}{2\sigma^2} - \frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2c_0\sigma^2} - \frac{b_0}{\sigma^2}\right) d\sigma^2$$

$$\propto \frac{\gamma((n+p)/2 + a_0)}{\left(\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta})}{2} + \frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2c_0} + b_0\right)^{(n+p)/2+a_0}}$$

$$\propto \left(\frac{(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{Y}-\boldsymbol{X\beta})}{2} + \frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2c_0} + b_0\right)^{-[(n+p)/2+a_0]}$$

$$\propto \left(1 + \frac{(\boldsymbol{\beta}-\boldsymbol{C})^T\boldsymbol{A}(\boldsymbol{\beta}-\boldsymbol{C})}{\boldsymbol{D}}\right)^{-[(n+2a_0+p)/2]}$$

$$\boldsymbol{A} = c_0\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{I}_p; \boldsymbol{C} = \boldsymbol{A}^{-1}\boldsymbol{\beta}_0 + c_0\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$\boldsymbol{D} = c_0\boldsymbol{Y}^T\boldsymbol{Y} + \boldsymbol{\beta}_0{}^T\boldsymbol{\beta}_0 + 2c_0b_0 - \boldsymbol{\beta}_0{}^T\boldsymbol{A}^{-1}\boldsymbol{\beta}_0 - c_0\boldsymbol{\beta}_0{}^T\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{Y} - c_0\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{\beta}_0 - c_0^2\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

So the marginal posterior distribution of $\boldsymbol{\beta}$ is non-central multivariate t distribution with degrees of freedom v $= n + 2a_0$, location parameter $\mathbf{Loc} = \boldsymbol{C}$, and scale parameter $\boldsymbol{\Sigma} = (\frac{(n+2a_0)\boldsymbol{A}}{\boldsymbol{D}})^{-1}$. Then we have both marginal and conditional distributions, $\pi(\boldsymbol{\beta}|data) \sim MVT(\text{v}, \mathbf{Loc}, \boldsymbol{\Sigma})$ and $\pi(\sigma^2|\boldsymbol{\beta}, data) \sim IG(a_1, b_1)$. The next step we will use Monte Carlo Method to draw many samples from these two distributions and estimated what we want to obtain.

3.2.2 Independent Prior and Gibbs Sampling

On the other hand, if we want to use MCMC in this estimation, we only need to draw full conditional distributions for both of them. Most of the procedures in calculations are similar to the Exact Sampling above and the main difference is $\boldsymbol{\beta}$ will follow a multivariate normal distribution instead of multivariate t distribution. Part of the procedures are shown as follows.

$$Posterior\ \pi(\boldsymbol{\beta}, \sigma^2|data)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{n/2+a_0+1} * \exp\left(-\frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2} - \frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2\tau_0^2} - \frac{b_0}{\sigma^2}\right)$$

$$\Longrightarrow \pi(\sigma^2|\boldsymbol{\beta}, data) \sim IG(n/2 + a_0, \frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{2} + b_0)$$

$$\pi(\boldsymbol{\beta}|\sigma^2, data) \propto \exp\left(-\frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2} - \frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2\tau_0^2}\right)$$

$$\propto \exp\left(-\frac{1}{2} * \frac{(\boldsymbol{\beta} - \sigma^2\boldsymbol{A}^{-1}\boldsymbol{\beta}_0 - \tau_0^2\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{Y})^T\boldsymbol{A}(\boldsymbol{\beta} - \sigma^2\boldsymbol{A}^{-1}\boldsymbol{\beta}_0 - \tau_0^2\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{Y})}{\sigma^2\tau_0^2}\right)$$

Where $\boldsymbol{A} = \sigma^2\boldsymbol{I}_p + \tau_0^2\boldsymbol{X}^T\boldsymbol{X}$. So as derived above the conditional distributions are multivariate normal and inverse gamma.

As we can see from the above derivation, there is a special property of normal distributions. Σ_d and Σ_0 are dispersion matric calculated for the parameters from the data and the prior respectively. Then we have the property as follows.

$$\text{Posterior Dispersion} = (\Sigma_d{}^{-1} + \Sigma_0{}^{-1})^{-1}$$

$$\text{Posterior Precision} = \text{Prior Precision} + \text{Data Precision}$$

$$\text{Precision} * \text{Mean in Post} = \text{Precision} * \text{Mean in Prior} + \text{Precision} * \text{Mean in Data}$$

3.2.3 Prediction using posterior samples

Similar to drawing new observations from posterior predictive distribution, we can make prediction on regression based on new predictor $\boldsymbol{x}^{new}$ and all previous data. For example, in linear regression $y = \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon; \varepsilon \sim N(0, \sigma^2)$.

$$f\left(y^{new} \middle| \boldsymbol{x}^{new}, (y_1, x_1), \dots, (y_n, x_n)\right)$$

$$= \int f(y^{new} | \boldsymbol{x}^{new}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}, \delta^2 | (y_1, x_1), \dots, (y_n, x_n))\, d\boldsymbol{\beta} d\sigma^2$$

In other words, given new observation, we calculate the regression value based on each parameter set from the simulation. And after getting the large amount of prediction results from all the parameters, summary the prediction.

Some kinds of regression models (for example, linear regression) have two types of prediction. One is mean prediction, and the other one is observation prediction. The former one does not include the error term and only calculate the regression value based on the regression mean. The 2nd prediction includes the error term, and after getting the predicted mean, will add the random error to the predicted value. Same liner regression example $y = \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon; \varepsilon \sim N(0, \sigma^2)$. Calculate

$y = \boldsymbol{x}^T\boldsymbol{\beta}$ for each parameter set, and we can get predicted mean. In the other hand, calculate $y = \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon$, and we can get predicted observation.

3.3 Regression with Binary Response

Now, we look at models with binary response so we have to use a suitable link function to relate the covariates to the response. We use the probit link here as it has a nice representation through auxiliary variable that works efficiently in a Monte Carlo method.

3.3.1 Probit Regression

Assume, Y is a binary variable can be 0 or 1, and $x_1, x_2, \ldots, x_p$ are covariates. In logistic regression we have

$$\mathrm{P}(\mathrm{Y} = 1) = \frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^T\boldsymbol{\beta})} = \mathrm{f}(W)$$

We have W=$\boldsymbol{x}^T\boldsymbol{\beta}$ (linear model and R → R), and $\mathrm{f}(W) = \frac{\exp(W)}{1+\exp(W)}$ (logistic and R→ (0,1)). In Probit Model instead of using f(W), we use F(x)=Φ(x) (CDF of standard normal distribution and R→ (0,1)). Then $\mathrm{P}(\mathrm{Y} = 1) = \Phi(\boldsymbol{x}^T\boldsymbol{\beta})$ and it is called Probit Model.

$$\mathrm{L}(\mathrm{data}|\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \mathrm{L}(y_i|\boldsymbol{\beta}) \propto \prod_{i=1}^{n} [\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{y_i} * [1 - \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{1-y_i}$$

$$\pi(\boldsymbol{\beta}) \propto \mathrm{MVN}_p(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$$

$$\pi(\boldsymbol{\beta}|\mathrm{data}) \propto \mathrm{L}(\mathrm{data}|\boldsymbol{\beta}) * \pi(\boldsymbol{\beta}) = \prod_{i=1}^{n} [\{[\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{y_i} * [1 - \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{1-y_i}] * \pi(\boldsymbol{\beta})$$

By introducing the cumulative probability function of standard normal, it is also extremely hard to simulate samples from this distribution no matter what method used here. To simplify this, latent variable (auxiliary variable) will be used. We introduce a latent variable z such that

$$\mathrm{y} = \begin{cases} 1, if\ z > 0 \\ 0, if\ z < 0 \end{cases}, and\ z \sim N(\mathbf{x}^T\boldsymbol{\beta}, 1)$$

Due to identifiability of the combination of the coefficients and the variance, we use constant 1 as the variance for z also in order to make sure that we have $\mathrm{P}(\mathrm{Y} = 1) = \Phi(\boldsymbol{x}^T\boldsymbol{\beta})$. By introducing z which is unobserved, y only directly depends on the value of z and also have the probability of $\mathrm{P}(\mathrm{Y} = 1) = \Phi(\boldsymbol{x}^T\boldsymbol{\beta})$.

$$P(Y = 1) = P(Z > 0) = P\left(\frac{Z - \boldsymbol{x}^T\boldsymbol{\beta}}{1} > \frac{-\boldsymbol{x}^T\boldsymbol{\beta}}{1}\right) = 1 - P\left(\frac{Z - \boldsymbol{x}^T\boldsymbol{\beta}}{1} \leq \frac{-\boldsymbol{x}^T\boldsymbol{\beta}}{1}\right)$$

$$= 1 - \Phi(-\boldsymbol{x}^T\boldsymbol{\beta}) = \Phi(\boldsymbol{x}^T\boldsymbol{\beta})$$

Given $\mathrm{y} = \begin{cases} 1, if\ z > 0 \\ 0, if\ z < 0 \end{cases}$, $z \sim N(\mathbf{x}^T\boldsymbol{\beta}, 1)$, $\pi(\beta) \propto \mathrm{MVN}_p(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$, and $\boldsymbol{\Sigma}_0 = \tau_0^2 * \boldsymbol{I}_p$. Then the joint posterior distribution and the posterior distribution of each parameter is calculated as follows.

$$\Pi(\boldsymbol{\beta}, \mathbf{Z}|\mathrm{data}) \propto \mathrm{L}(\mathrm{data}|\mathbf{Z}) * \pi(\mathbf{Z}|\boldsymbol{\beta}) * \pi(\boldsymbol{\beta})$$

$$\propto \left\{\prod_{i=1}^{n}\left[1(z_i > 0)^{y_i} * 1(z_i < 0)^{1-y_i} * \exp\left(-\frac{(z_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2}\right)\right]\right\} * \exp(-\frac{(\boldsymbol{\beta} - \boldsymbol{\beta_0})^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta_0})}{2})$$

$$\propto \left\{\prod_{i=1}^{n}\left[1(z_i > 0)^{y_i} * 1(z_i < 0)^{1-y_i} * \exp\left(-\frac{(z_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2}\right)\right]\right\} * \exp(-\frac{(\boldsymbol{\beta} - \boldsymbol{\beta_0})^T(\boldsymbol{\beta} - \boldsymbol{\beta_0})}{2{\tau_0}^2})$$

$$\Pi(\boldsymbol{\beta}|\mathbf{Z}, \mathrm{data}) \propto \exp(-\frac{(\boldsymbol{\beta} - \boldsymbol{A^{-1}\beta_0} - {\tau_0}^2\boldsymbol{A^{-1}X}^T\mathbf{Z})^T\boldsymbol{A}(\boldsymbol{\beta} - \boldsymbol{A^{-1}\beta_0} - {\tau_0}^2\boldsymbol{A^{-1}X}^T\mathbf{Z})}{2{\tau_0}^2})$$

Where $\boldsymbol{A} = \boldsymbol{I}_p + {\tau_0}^2\boldsymbol{X}^T\boldsymbol{X}$**.** Initial value of β_0 is obtained as follows.

$$\mathrm{P}(y = 1) = \Phi(\boldsymbol{x}^T\boldsymbol{\beta}) = \Phi\big(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\big); set\ x_i = 0\ \forall i$$

$$\mathrm{P}(y=1)=\Phi(\beta_0)\Longrightarrow\beta_0=\Phi(p_0)^{-1}$$

Other part of the initial value of $\boldsymbol{\beta}$ will be calculated from the least square method. Then we can derive that $\boldsymbol{\beta}\sim MVN_p(\boldsymbol{A^{-1}\beta_0}+{\tau_0}^2\boldsymbol{A^{-1}X^T\mathbf{Z}},(\frac{\boldsymbol{A}}{{\tau_0}^2})^{\boldsymbol{-1}})$.

$$\Pi(z_i|\boldsymbol{\beta},y_i=1)\propto 1(z_i>0)*\exp\left(-\frac{(z_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2}\right)$$

$$\Pi(z_i|\boldsymbol{\beta},y_i=0)\propto 1(z_i<0)*\exp\left(-\frac{(z_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2}\right)$$

$z_i\sim Truncated\ N(\boldsymbol{x}_i^T\boldsymbol{\beta},1)$. And then we can use formula as follows to simulate z_i from the truncated normal distribution.

$$\mu\sim\mathrm{unif}(0,1);\mathrm{X}\sim\text{Truncated N}(\mu,\sigma^2)\ within\ (a,b)$$

$$\mathrm{P}(\mathrm{X}\le\mathrm{c})=\frac{\Phi\left(\frac{c-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}$$

After simulating a large amount of parameters, in terms of a new data point $\boldsymbol{x}^{new}$, making a prediction on the category of y is our next step.

$$\mathrm{P}(y^{new}=1)=\int\mathrm{P}(y^{new}=1|z^{new})*f(z^{new})\,dz^{new}=\Phi\left(\boldsymbol{x}^{newT}\boldsymbol{\beta}\right)$$

$$\Phi\left(\boldsymbol{x}^{newT}\boldsymbol{\beta}_1\right)=p_1,\Phi\left(\boldsymbol{x}^{newT}\boldsymbol{\beta}_2\right)=p_2,\dots,\Phi\left(\boldsymbol{x}^{newT}\boldsymbol{\beta}_N\right)=p_N;\bar{p}=\frac{\sum_{i=1}^{N}p_i}{N}$$

Or we can use hierarchical method here.

$$z_1^{new}\sim N\left(\boldsymbol{x}^{newT}\boldsymbol{\beta}_1,1\right),z_2^{new}\sim N\left(\boldsymbol{x}^{newT}\boldsymbol{\beta}_2,1\right),\dots,z_N^{new}\sim N\left(\boldsymbol{x}^{newT}\boldsymbol{\beta}_N,1\right)$$

$$\{y_1^{new},y_2^{new},\dots,y_N^{new}\};\mathrm{p}(y^{new}=1)=\frac{\#(\,y^{new}=1)}{N}$$

3.3.2 Ordinal Probit regression

Assume Y has k ordinal categories, and we change categories into $y = 1,2, \dots, k$. We will use (k+1) α's ($\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_{k+1}$, with k-1 unknown parameters) to separate the real number line into k ordinal categories with $\alpha_1 = -\infty$ and $\alpha_{k+1} = +\infty$. To make this sampling possible and simpler, we will also introduce a latent variable Z such that y=i if $z \in (\alpha_i, \alpha_{i+1})$ and $Z \sim N(\boldsymbol{x}^T\boldsymbol{\beta}, \sigma^2)$. Then the response variable Y is only directly determined by Z. Here p_i represents the probability y will fall into the i^{th} category. P's need to satisfy $\sum_{i=1}^{k} p_i = 1$ with (k-1) degree of freedom. And then we will have

$$p_i = P(y = i) = P(\alpha_i < z < \alpha_{i+1}) = P(z < \alpha_{i+1}) - P(z < \alpha_i)$$

$$= \Phi\left(\frac{\alpha_{i+1} - \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right) - \Phi(\frac{\alpha_i - \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma})$$

For that these two sets $(\alpha_i = 2, \alpha_{i+1} = 4, \boldsymbol{\beta} = (2,3)^T, \sigma = 7)^T$ and $(\alpha_i = 20, \alpha_{i+1} = 40, \boldsymbol{\beta} = (20,30)^T, \sigma = 70)^T$ will give out the same probability p_i, we need to fix one of these parameters, which is generally $\sigma = 1$. Then we will have the probability as

$$p_i = \Phi(\alpha_{i+1} - \boldsymbol{x}^T\boldsymbol{\beta}) - \Phi(\alpha_i - \boldsymbol{x}^T\boldsymbol{\beta})$$

Free parameters are $\{\alpha_2, \alpha_3, \dots, \alpha_k, \beta_0, \beta_1, \dots, \beta_p\}$. For any constant c if we have $\{\alpha_2 + c, \alpha_3 + c, \dots, \alpha_k + c, \beta_0 + c, \beta_1, \dots, \beta_p\}$, then

$$\alpha_2 - \boldsymbol{x}^T\boldsymbol{\beta} = \alpha_2 - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p = (\alpha_2 + c) - (\beta_0 + c) - \beta_1 x_1 - \cdots - \beta_p x_p$$

The parameterization is not identifiable. And generally people will tend to set $\alpha_2 = 0$ to keep it identifiable. So we have (k-2) free α's. Similar to Probit Model, independent multivariate normal prior will be used for $\boldsymbol{\beta}$ ($\boldsymbol{\beta} \sim MVN_p(\boldsymbol{\beta}_0, \tau_0^2 * \boldsymbol{I}_p)$) and univariate normal prior will be used for Z. Prior for α's is shown as follows. And because α's only depend on the value of Z, I will derive the posterior distribution for α's directly.

$$\pi(\alpha_2, \alpha_3, \dots, \alpha_k) \propto 1(\alpha_k > \alpha_{k-1} > \cdots > \alpha_3 > 0)$$

$$y_j = 2 \; if \;\; \alpha_2 = 0 < z_j < \alpha_3; y_j = 3 \; if \;\; \alpha_3 < z_j < \alpha_4; \dots$$

$$y_j = k \; if \;\; \alpha_k < z_j < \alpha_{k+1} = \infty$$

$$\Longrightarrow \max_{y_j=i-1} z_j \le \alpha_i \le \max_{y_j=i} z_j \Longrightarrow \pi(\alpha_i | z, data) \sim unif\left(\max_{y_j=i-1} z_j, \max_{y_j=i} z_j\right)$$

Jointly posterior distribution and posterior distribution for other parameters are derived as follows.

$$\pi(\boldsymbol{\alpha}, Z, \boldsymbol{\beta} | data) \propto \prod \prod \left(\mathrm{L}\left(y_j | z_j, \boldsymbol{\alpha}\right) \pi\left(z_j | \boldsymbol{\beta}\right)\right) * \pi(\boldsymbol{\alpha}) * \pi(\boldsymbol{\beta})$$

$$\propto \prod_{j=1}^{n} \{1\left(\alpha_i < z_j < \alpha_{i+1}\right)^{1(y_j=i)} * \exp\left(-\frac{\left(z_j - \boldsymbol{x}_j^T \boldsymbol{\beta}\right)^2}{2}\right)\}$$

$$* 1(\alpha_k > \alpha_{k-1} > \cdots > \alpha_3 > 0) * \exp(-\frac{(\boldsymbol{\beta} - \boldsymbol{\beta_0})^T (\boldsymbol{\beta} - \boldsymbol{\beta_0})}{2{\tau_0}^2})$$

$$\boldsymbol{\beta} \sim MVN_p\left(\boldsymbol{A^{-1}\beta_0} + {\tau_0}^2 \boldsymbol{A^{-1}X^T Z}, \left(\frac{\boldsymbol{A}}{{\tau_0}^2}\right)^{-\boldsymbol{1}}\right); \boldsymbol{A} = \boldsymbol{I}_p + {\tau_0}^2 \boldsymbol{X^T X}$$

$$\Pi\left(z_j | \boldsymbol{\beta}, y_j = \mathrm{i}\right) \propto 1\left(\alpha_i < z_j < \alpha_{i+1}\right)^{1(y_j=i)} * \exp\left(-\frac{\left(z_j - \boldsymbol{x}_j^T \boldsymbol{\beta}\right)^2}{2}\right)$$

z_j follows truncated normal with pdf above. Procedures to get the initial value for β_0 and free α's are shown as follows.

$$\widehat{p_1} = P(y \le 1) = \Phi(0 - \beta_0); \; \widehat{p_2} = P(y \le 2) = \Phi(\alpha_3 - \beta_0); \dots$$

Then we can simulate $\boldsymbol{\alpha}, Z, \boldsymbol{\beta}$ from the posterior distribution derived above sequentially.

3.4 Poisson Regression

Data $\{y_1, y_2, y_3 \dots, y_n\}$ are counts and we want to fit a regression model on the observations with some covariates $\{x_{i1}, x_{i2}, \dots, x_{ip}\}(for \; i = 1, \dots, n)$. Assume count data follows Poisson

distribution, in other words $y_i \sim Poisson(\lambda_i)$ with some λ_i's. And we need to regression on λ_i with covariates for each observation. Because λ_i only can take positive real numbers, we do some transformation on the λ_i.

$$\log(\lambda_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \ ; \ (\boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_i) \rightarrow \lambda_i \rightarrow y_i$$

To make the regression easy and can be realized, we do another transformation here $\log(\lambda_i) = \eta_i$ (because η_i can be any real number and λ_i only can be positive). Then the relationship will become as follows.

$$e^{\eta_i} = \lambda_i; \ y_i \sim Poisson(e^{\eta_i})$$

And due to the above relationship, initial values of η_i in MCMC is as follows.

$$\eta_i = \log(y_i + 0.5)$$

I will do MH sampling and Slice sampling to realize the MCMC method to do the parameters' estimation as follows.

3.4.1 Using Metropolis-Hastings within Gibbs

In MH sampling, for jointly binary posterior distribution of $\boldsymbol{\beta}$ and σ^2 I will do Exact sampling within MCMC. So here I will use dependent prior for $\boldsymbol{\beta}$, which is $\boldsymbol{\beta} \sim MVN_p(\boldsymbol{\beta}_0, c_0\sigma^2)$ (multivariate normal distribution). Similar to linear regression, inverse gamma prior will be used for σ^2, ($\sigma^2 \sim IG(a_0, b_0)$). Proposed distribution (kernel) for η is normal distribution with mean equal to the most recent η and constant being some suitable constant. In other words, $\eta^{new} \sim N(\eta^{old}, \tau^2)$. Next I derive the posterior distribution for each parameter and illustrate the MH procedure.

$$\text{Likelihood} = \prod_{i=1}^{n} \frac{1}{y_i!} (e^{\eta_i})^{y_i} e^{(-e^{\eta_i})} \propto \prod_{i=1}^{n} e^{\eta_i y_i} e^{(-e^{\eta_i})}$$

$$\pi(\eta_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2}\frac{(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{\sigma^2}) \propto \frac{1}{\sqrt{\sigma^2}}\exp(-\frac{1}{2}\frac{(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{\sigma^2})$$

$$\pi(\boldsymbol{\beta}) \propto \frac{1}{(c_0\sigma^2)^{p/2}}\exp\left(-\frac{1}{2}\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{c_0\sigma^2}\right);\ \pi(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{a_0+1}\exp(-\frac{b_0}{\sigma^2})$$

$$\text{Posterior} \propto \text{Likelihood} * \pi(\eta_i) * \pi(\boldsymbol{\beta}) * \pi(\sigma^2)$$

And similar to the Exact sampling for the linear regression, we know that posterior distribution of σ^2 is inverse gamma distribution and the marginal posterior distribution of $\boldsymbol{\beta}$ is non-central multivariate t distribution.

$$\sigma^2 \sim \text{IG}(\frac{n+p}{2} + a_0, \frac{1}{2}\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{c_0} + \frac{(\eta - \boldsymbol{X\beta})^T(\eta - \boldsymbol{X\beta})}{2} + b_0)$$

$$\boldsymbol{\beta} \sim \text{MVT}(\boldsymbol{\Sigma}, \boldsymbol{\mu}, \text{v}); \text{v} = \text{n} + 2a_0, \boldsymbol{\mu} = \boldsymbol{A}^{-1}\boldsymbol{\beta}_0 + c_0\boldsymbol{A}^{-1}\boldsymbol{X}^T\eta, \boldsymbol{\Sigma} = [\frac{\boldsymbol{A}(\text{n} + 2a_0)}{\boldsymbol{C}}]^{-1}$$

$$with\ \boldsymbol{C} = c_0\eta^T\eta + \boldsymbol{\beta}_0{}^T\boldsymbol{\beta}_0 + 2c_0b_0 - \boldsymbol{\beta}_0{}^T\boldsymbol{A}^{-1}\boldsymbol{\beta}_0 - c_0\boldsymbol{\beta}_0{}^T\boldsymbol{A}^{-1}\boldsymbol{X}^T\eta - c_0\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{\beta}_0$$
$$- c_0^2\eta^T\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^T\eta\ ; \text{and}\ \ \boldsymbol{A} = c_0\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{I}_p$$

$$\text{Posterior}\ \eta_i\ \propto e^{\eta_i y_i} * e^{(-e^{\eta_i})} * \exp\left(-\frac{1}{2}\frac{(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{\sigma^2}\right) = f(\eta_i)$$

We will use MH sampling here for η_i. With normal proposal and target $f(\eta_i)$. And due to the normal symmetry the acceptance probability is simplified as follows, which is similar to MH in univariate normal simulation.

$$p_{\eta^{old}\to\eta^{new}} = \frac{f(\eta^{new})}{f(\eta^{old})} * \frac{q(\eta^{old}|\eta^{new})}{q(\eta^{new}|\eta^{old})} = \frac{f(\eta^{new})}{f(\eta^{old})}$$

And we propose a new η^{new} from normal distribution with mean η^{old} and some variance, and we calculated the acceptance probability to compare with a uniform random number within (0, 1). And in the real simulation I take the logarithm of the probability to reduce the complicated calculation. The acceptance will be tracked and I will change the variance of the proposal to keep

it within 30%-40%. Then iteratively do the simulation for all the parameters until get a relatively large sample.

3.4.2 Using slice sampling within Gibbs

Instead of using MH sampling to simulate η, I will use Slice sampling to simulate η. Priors used here is exactly same with the priors in MH sampling just mentioned above. But there is no proposal, and I will draw samples of η by adding a random variable u to implement Slice sampling. I will only explain the sampling of η in details next.

$$\text{Posterior } \eta_i \ \propto \underbrace{e^{(-e^{\eta_i})}}_{h(\eta_i)} * \underbrace{e^{\eta_i y_i} * \exp\left(-\frac{1}{2}\frac{(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{\sigma^2}\right)}_{g(\eta_i)}$$

$$\pi(u_i|\eta_i) \sim \text{unif}\left(0, e^{(-e^{\eta_i})}\right); \pi(u_i|\eta_i) \propto \frac{1}{e^{(-e^{\eta_i})}} * 1(u_i < e^{(-e^{\eta_i})})$$

$$f(u_i, \eta_i) \propto e^{\eta_i y_i} * \exp\left(-\frac{1}{2}\frac{(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{\sigma^2}\right) * 1(u_i < e^{(-e^{\eta_i})})$$

$$f(\eta_i|u_i) \propto e^{\eta_i y_i} * \exp\left(-\frac{1}{2}\frac{(\eta_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{\sigma^2}\right) * 1(\eta_i < \log(-\log(u_i)))$$

$$f(\eta_i|u_i) \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta} + \sigma^2 y_i, \sigma^2) \ \ ; \ \eta_i < \log(-\log(u_i))$$

So we can use full conditional distribution above to draw samples in turn.

3.5 First order Autoregressive Time Series

In economy, marketing and many other areas, data flows associated with time are very common and important. And most of the time, time series data like this tends to have autocorrelation between themselves. In other words, the future data value is affected by the previous data values and can be predicted based on the previous data values. For these kinds of time series data

$x_1, x_2, \dots, x_T$ $(t = 1,2, \dots, T)$, we want to fit a model which can capture the relationships between themselves as time interval changing and make a prediction on it. As the beginning of the it, I first fit an Autoregressive model with one lag, AR(1). The model is as follows

$$x_t - \mu = \phi(x_{t-1} - \mu) + \varepsilon_t; \varepsilon_t \sim N(0, \sigma^2)$$

Next I will use MCMC simulation within Bayesian to do the estimation of the parameters. In AR(1) model, time series needs to be stationary, in other words constant mean, constant variance and constant covariance between same time interval. Also to keep the time series, we need the modulus of ϕ strictly less than 1, which can satisfy that all the roots of the characteristic equation are great than 1. Then I will use uniform distribution within (-1,1) as the prior of ϕ. For μ and σ^2, I will use the independent prior distributions, which are normal distribution and inverse gamma distribution respectively. There is another parameter of the time series model which needs to be estimated, the very beginning of the time series x_0 to get x_1. To get the distribution of x_0, and keep all the time series have the same variance, I will use a normal distribution with mean and variance as a function of other parameters.

$$\text{Likelihood} \propto \prod_{t=1}^{T} (\frac{1}{\sigma^2})^{\frac{1}{2}} \exp(-\frac{1}{2}\frac{(X_t - \mu - \Phi * (X_{t-1} - \mu))^2}{\sigma^2})$$

$$\text{Priors}: \Phi \sim \text{Unif}(-1,1)\ ;\ \sigma^2 \sim \text{IG}(a_0, b_0)\ ;\ x_0 \sim N\left(\mu, \frac{\sigma^2}{1-\Phi^2}\right)\ ;\ \mu \sim N(\mu_0, \sigma_0^2)$$

$$\Longrightarrow \pi(x_0|\sigma^2, \mu, \Phi) \sim N(\mu + x_1\Phi - \mu\Phi, \sigma^2)$$

$$\Longrightarrow \pi(\sigma^2|\mu, x_0, \Phi) \sim IG(a_1, b_1)$$

$$a_1 = \frac{1+T}{2} + a_0\ ;\ b_1 = \frac{1}{2}\sum_{t=1}^{T}(x_t - \mu - \Phi(x_{t-1} - \mu))^2 + \frac{1}{2}(1-\Phi^2)(x_0 - \mu)^2 + b_0$$

$$\Longrightarrow \pi(\mu|\sigma^2, x_0, \Phi) \sim N(\mu_1, \sigma_1^2)$$

$$\mu_1 = \frac{[\sigma_0^2(1-\Phi)\sum(x_t - \Phi x_{t-1}) + \sigma_0^2 x_0(1-\Phi^2) + \sigma^2\mu_0]}{A}; \sigma_1^2 = \frac{\sigma_0^2\sigma_1^2}{A}$$

$$A = \sigma_0^2 T(1-\Phi)^2 + \sigma_0^2(1-\Phi^2) + \sigma^2$$

$$\Phi|\sigma^2, \mu, x_0 \propto N(\mu_2, \sigma_2^2)1(-1 < \Phi < 1)\sqrt{1-\Phi^2}\exp(\frac{\Phi^2(x_0-\mu)^2}{2\sigma^2})$$

$$\mu_2 = \frac{\sum(x_t - \mu)(x_{t-1} - \mu)}{\sum(x_{t-1} - \mu)^2}; \sigma_2^2 = \frac{\sigma^2}{\sum(x_{t-1} - \mu)^2}$$

We can find that the posterior distribution of Φ is not standard distribution, so I will use Metropolis Hastings Sampling with the proposed distribution is independent. Proposal is the first part of the posterior, which is $N(\mu_2, \sigma_2^2)$ truncated within (-1,1). Then the accept probability is

$$Prob = \frac{\sqrt{1-\Phi_{new}^2} * \exp(\frac{\Phi_{new}^2(x_0-\mu)^2}{2\sigma^2})}{\sqrt{1-\Phi_{old}^2} * \exp(\frac{\Phi_{old}^2(x_0-\mu)^2}{2\sigma^2})}$$

Based on the AR(1) model estimated above, prediction of the time series is our next step. So I derive the posterior predictive distributions as follows.

$$x_{T+1} - \mu = \Phi(x_T - \mu) + N(0, \sigma^2) \Longrightarrow x_{T+1} = \mu + \Phi(x_T - \mu) + N(0, \sigma^2)$$

$$x_{T+2} = \mu + \Phi(x_{T+1} - \mu) + N(0, \sigma^2) = \mu + \Phi\big(\Phi(x_T - \mu) + N(0, \sigma^2)\big) + N(0, \sigma^2)$$

$$= \mu + \Phi^2(x_T - \mu) + \Phi N(0, \sigma^2) + N(0, \sigma^2)$$

$$\Longrightarrow x_{T+N} = \mu + \Phi^N(x_T - \mu) + \mathbf{y}'\boldsymbol{z}\ ;\ \mathbf{y} = (\Phi^0, \Phi^1, \dots, \Phi^{N-1})'\ ;\ \boldsymbol{z} = unif(N)$$

From the posterior predictive distributions above, we can find that as predictive time interval increasing the variance is increasing. This is also means that in the long term prediction, the uncertainty is much larger than the short term prediction. This is proved as follows.

$$Var(x_{T+1}|x_T) = \sigma^2\ ;\ Var(x_{T+2}|x_T) = \Phi^2\sigma^2 + \sigma^2\ ;$$

$$Var(x_{T+N}|x_T) = \sigma^2 \sum_{j=0}^{N-1} \Phi^{2j}$$

3.6 Data Analysis

In this section, we use several real-world datasets to carry out parameter estimation and out-of-sample prediction.

3.6.1 Birth-rate Dataset

This dataset (Weintraub 1962) consists of Birth Rates, per capita income, proportion of population in farming and infant mortality during early 1950s for 30 nations. I use Exact Sampling with the regression function $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon\ ;\ \varepsilon \sim N(0, \sigma^2)$, which is BR~PCI+PDF with both numeric covariates from data set. Before regression procedure, I leave out Philippines and Austria to check the prediction accuracy. Figure 8 and Table 6 give out the simulation result and posterior distribution summaries for variance of normal error and all the coefficients of covariates.

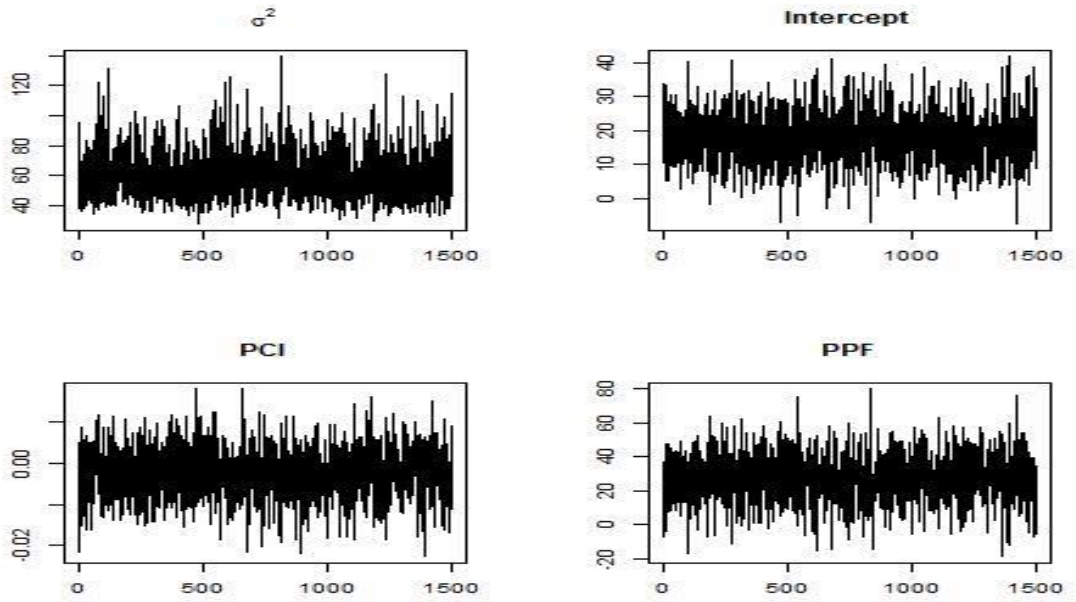

Figure 8. Parameters simulation from exact MC in linear regression

Parameter	Mean	95% Credible Set
σ^2	59.54153	(0, 89.15135)
Intercept	17.9496	(4.30476, 32.72799)
PCI	-0.002356502	(-0.014615, 0.008894)
PPF	27.20564	(0.224339, 53.159042)

Table 6 Posterior summary for exact MC in linear regression

After finishing the regression estimation, I want to check the regression accuracy of this model, and plug in the Philippines and Austria data. As have explained previous, I make both mean and observation prediction. The simulation of the posterior predictive distribution and the summary of the predictive distribution are given in Figure 9 and Table 7. It shows that the mean prediction credible interval, which incorporates smaller variance and has a smaller interval, does not give a good prediction. However, the observation prediction interval, which has more variance and larger, include the true value.

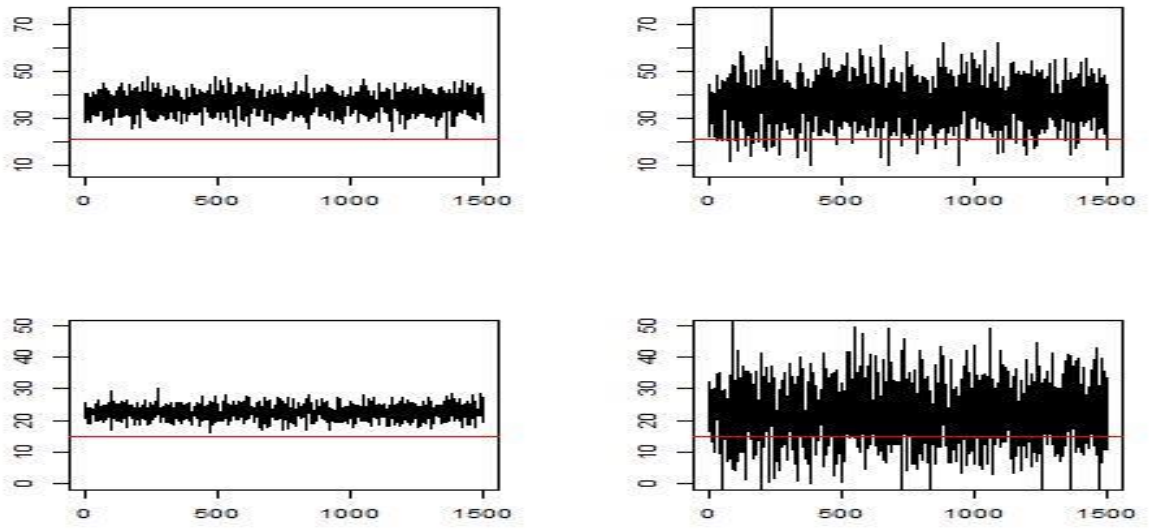

Figure 9. Prediction from exact MC in linear regression

Red line in Figure 9 is the true value of the prediction data.

	True Obs	95% Credible Mean	95% Credible Obs
Philippines	21.3	(29.22596, 43.59412)	(18.92557, 52.87955)
Austria	14.8	(18.89994, 26.60434)	(6.193693, 37.564797)

Table 7 Prediction results from exact MC in linear regression

After that, I use MCMC Sampling with the regression function $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$; $\varepsilon \sim N(0, \sigma^2)$, which is IMR~PCI+PDF with both the same numeric variables. I also leave out Philippines and Austria to check the prediction accuracy. Similar to the Exact example above, Figure 10 and Table 8 gives out the simulation and summaries of the parameters.

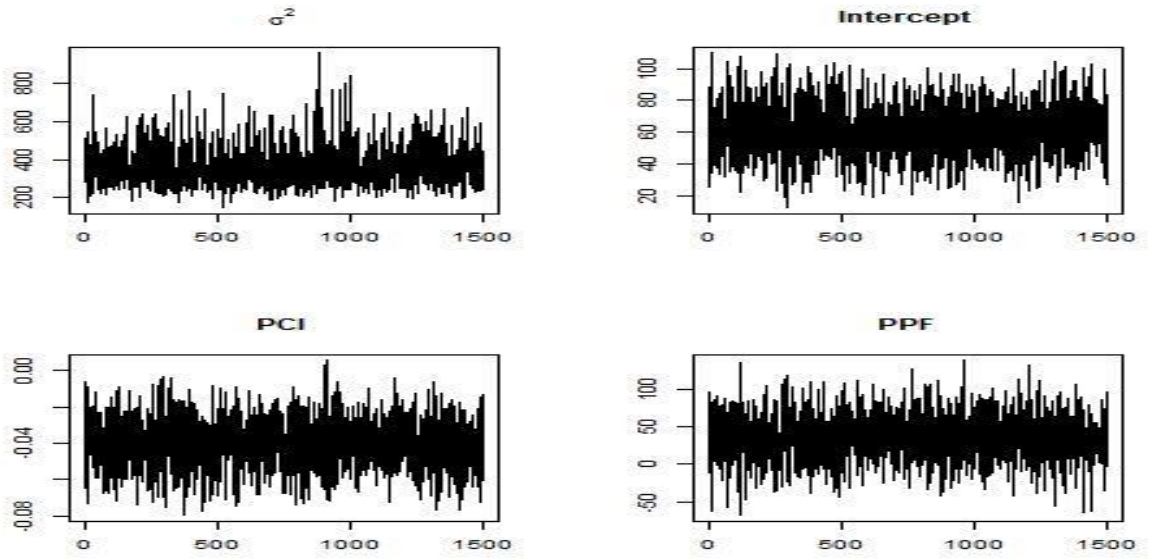

Figure 10. Parameters simulation from MCMC in linear regression

Parameter	Mean	95% Credible Set
σ^2	360.3697	(0, 555.2955)
Intercept	61.19677	(25.38162, 93.97517)
PCI	-0.04039937	(-0.06741, -0.01241)
PPF	36.49901	(-26.55976, 103.13510)

Table 8 Posterior summary for MCMC in linear regression

After I have simulated large amount of sets of parameters from the posterior distribution, I use the same prediction method for both mean prediction and observation prediction. As we can see from Figure 11, the simulation results from posterior predictive distribution of this model is worse than the last one which parameters are estimated through Exact sampling. And from Table 9, we also can find in this prediction, only the observation credible interval for Austria includes the true value. But for Philippines, neither observation credible interval nor mean credible interval includes the true value.

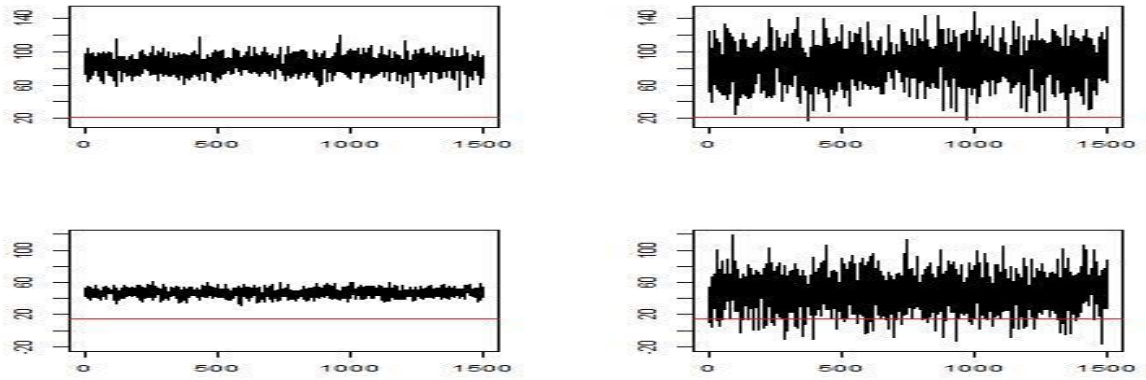

Figure 11. Prediction from MCMC in linear regression

	True Obs	95% Credible Mean	95% Credible Obs
Philippines	21.3	(66.82059,102.41722)	(41.55966, 126.07538)
Austria	14.8	(37.43068, 55.58567)	(8.510534, 85.203908)

Table 9 Prediction results from MCMC in linear regression

3.6.2 Low Birth Weight Data

This dataset (Hosmer et al. 2013) includes information on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. Data were collected at Baystate Medical Center, Springfield, Massachusetts during 1986. In the dataset, I choose binary variable LOW as response with two categories (0 and 1). Predictors are AGE (numeric), LWT (numeric), RACE (categorical with 3 categories 1, 2, and 3), SMOKE (categorical with 2 categories 0 and 1), PTL (categorical with 4 categories 0, 1, 2, and 3), HT (categorical with 2 categories 0 and 1), UI (categorical with 2 categories 0 and 1), and FTV (categorical with 6 categories 0, 1, 2, 3, 4, and 6). After check the data, find that there is only one observation fall in category 3 of variable PTL and there is only one observation fall in category 6 of variable FTV. We want to create dummy variables for categorical predictors. When you create one column for each category of a predictor, the underlying assumption is that you have a reasonable number of observations falling into that category. So for predictor PTL, I will combine categories 3 and 2 together. Similarly, for predictor FTV I will combine categories 4 and 6. Then the model and the parameters which we need to estimated is as follows.

$$\text{LOW} = \begin{cases} 1, if\ z > 0 \\ 0, if\ z < 0 \end{cases}, and\ z \sim N(\mathbf{x}^T \boldsymbol{\beta}, 1)$$

So there are totally 13 coefficients β's (excluding intercept). Then I will use the formula and method illustrated above to do the MCMC and get the simulation of all the parameters. Figure 12 shows the simulation of four coefficients.

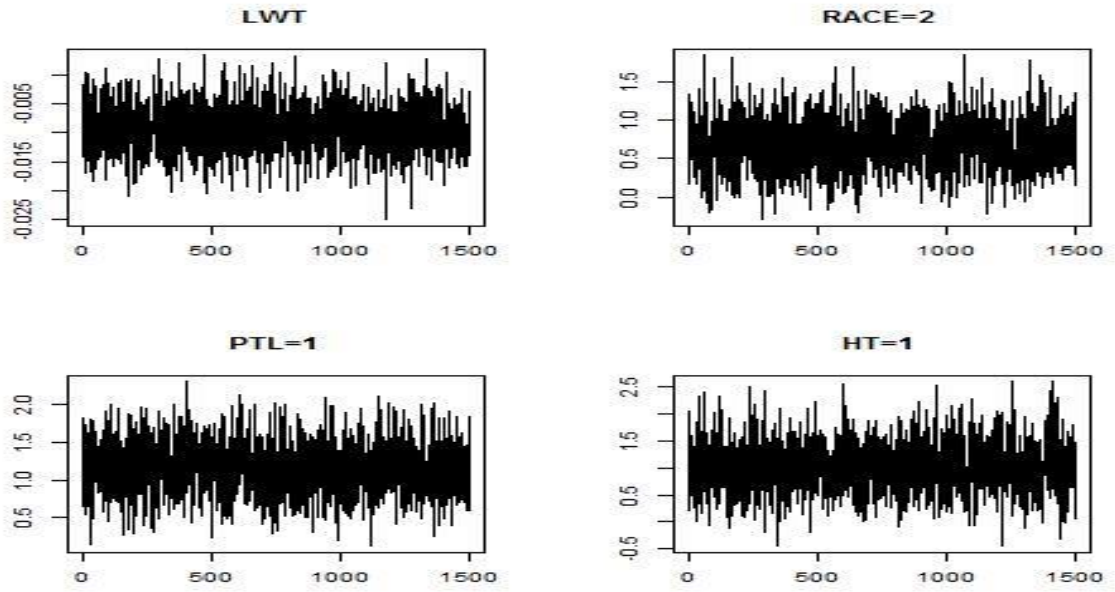

Figure 12. Parameters simulation in Probit regression

And when I am doing the regression, I leave out two observations to check the prediction accuracy. And the prediction probability is shown as below in Figure 13 and the Table 10 gives out the mean and 95% credible interval of the prediction probabilities. And from the prediction results, observation 1 has the mean probability 0.2513769 to be in category 1 which is a good prediction with true category 0. Meanwhile, observation 2 also has a large probability to be in category 1, which also gives out a reasonable prediction.

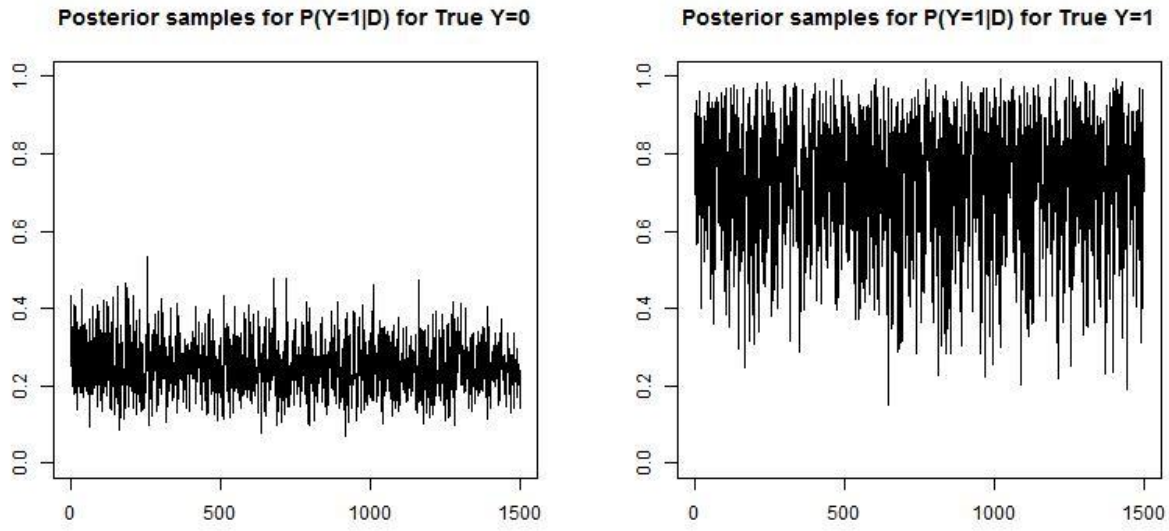

Figure 13. Prediction in Probit regression

	True Cat	Mean Prob.	95% Credible Interval
Observation 1	0	0.2513769	(0.1341832, 0.3963629)
Observation 2	1	0.7369638	(0.3590383, 0.9684874)

Table 10 Prediction results in Probit regression

3.6.3 Copenhagen Housing Condition Dataset

This dataset (Madsen 1976) classifies 1681 residents of twelve areas in Copenhagen, Denmark in terms of: (i) the type of housing they had (1=tower blocks, 2=apartments, 3=atrium houses and 4=terraced houses), (ii) their feeling of influence on apartment management (1=low, 2=medium,3=high), (iii) their degree of contact with neighbors (1=low, 2=high), and (iv) their satisfaction with housing conditions (1=low, 2=medium, 3=high). In this example, I use satisfaction with 3 ordinal categories (low, medium, and high) as response, and I use housing (tower, apartments, atrium, and terraced four categories), influence (low, medium, and high 3

categories), and contact (low and high 2 categories) as predictor, which are all categorical variables. I will use the method illustrated in the Ordinal Probit regression section. Here k=3, $\boldsymbol{\beta}$ is three dimensional vector, and one free α needs to be estimated. But we need to construct dummy variables to implement the regression on the categorical predictors. That is why we have 6 coefficients (excluding intercept) to estimate.

$$P(y = i) = \Phi\left(\frac{\alpha_{i+1} - \boldsymbol{x}^T\boldsymbol{\beta}}{1}\right) - \Phi(\frac{\alpha_i - \boldsymbol{x}^T\boldsymbol{\beta}}{1})$$

Simulated parameters and their posterior summaries are shown in Figure 14 and Table 11.

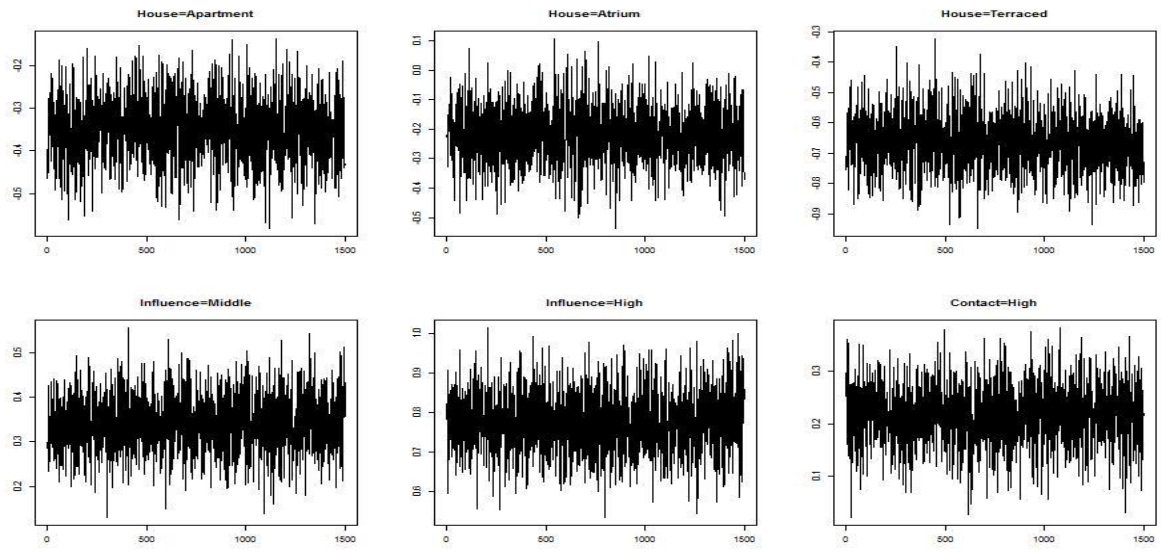

Figure 14. Parameters simulation in ordinal Probit regression

Covariate Effect	Mean	95% Credible Set
House: Apartment	-0.3518009	(-0.4946149, -0.2112778)
House: Atrium	-0.2178059	(-0.40814113, -0.04152039)
House: Terraced	-0.6668755	(-0.8447894, -0.4856407)
Influence: Middle	0.3438438	(0.2214866, 0.4651268)
Influence: High	0.7775783	(0.6363497, 0.9259575)
Contact: High	0.220256	(0.1116209, 0.3330477)

Table 11 Posterior summary in ordinal Probit regression

I also leave three data points which belong to 1th, 2nd, and 3rd categories respectively out to make the make prediction on them and make comparison with the real results. Posterior prediction results are shown in Table 12 as follows.

	True Cat	Cat1 Pred Prob	Cat2 Pred Prob	Cat3 Pred Prob
Observation1	1	0.37852361	0.2850438	0.3364326
Observation2	2	0.25706374	0.2745559	0.4683803
Observation3	3	0.09635675	0.1872627	0.7163805

Table 12 Prediction results in ordinal Probit regression

From the prediction results in Table 12, we can find that for observation 1 and observation 3 the model gives out the right prediction. But for observation 2 the model gives out the wrong prediction as category 3.

3.6.4 Ear Infection in Swimmers Dataset

This dataset (Hand et al. 1994) come from the 1990 Pilot Surf/Health Study of New South Wales Water Board, Sydney, Australia. The first column takes values 1 or 2 according to the recruit's

perception of whether (s)he is a Frequent Ocean Swimmer, the second column has values 1 or 4 according to recruit's usually chosen swimming location (1 for non-beach, 4 for beach), the third column has values 2 (aged 15-19), 3 (aged 20-25), or 4 (aged 25-29), the fourth column has values 1 (male) or 2 (female) and finally, the fifth column has the number of self-diagnosed ear infections that were reported by the recruit. For analyzing this dataset, I will use count data as response, and use 4 categorical predictors which result in the 5 dummy column in X matrix (excluding intercept). We have $y_i \sim Poisson(e^{\eta_i})$ and $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$. First, I use Exact sampling within MCMC including MH sampling with normal proposal $\eta^{new} \sim N(\eta^{old}, \tau^2)$ to do the simulation. Before that, I choose the acceptance rate which produces reasonable acceptance rate within 30% - 40%. Then, I present the simulated parameters and summary of the parameters in Figure 15 and Table 13 as follows.

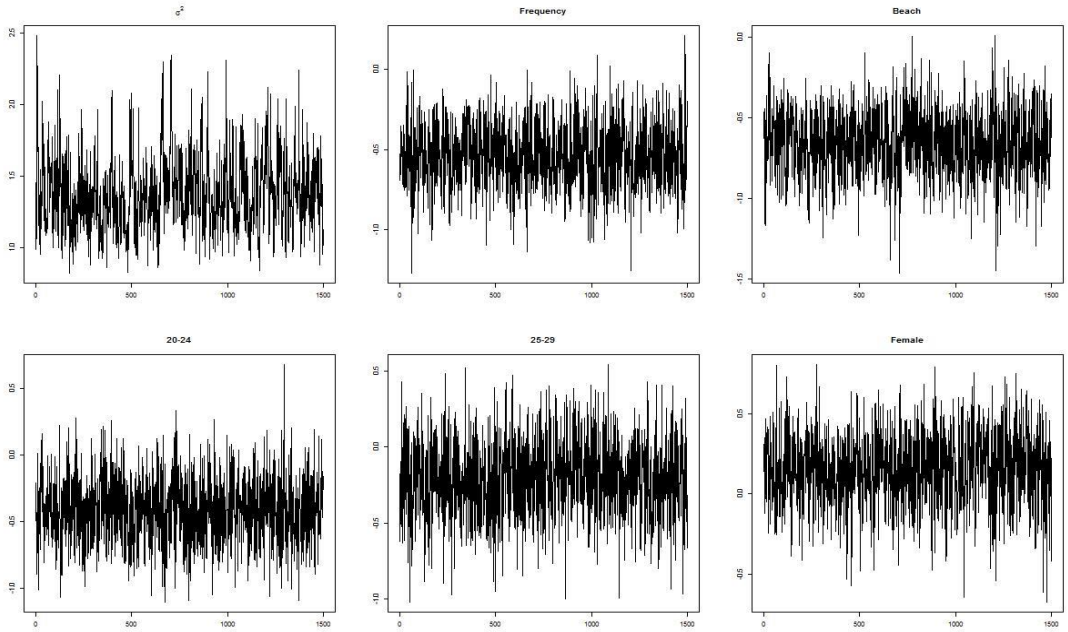

Figure 15. Parameters Simulation (MH) in Poisson regression

Parameter	Mean	95% Credible Set
σ^2	1.378531	(0, 1.828749)
Frequent	-0.5510663	(-0.9303471, -0.1701346)
Beach	-0.6616766	(-1.0670469, -0.2718842)
20-24	-0.412349	(-0.88021871, 0.05270854)
25-29	-0.1993489	(-0.6889213, 0.2801244)
Female	0.1522128	(-0.2809000, 0.5737834)

Table 13 Posterior summary (MH) in Poisson regression

Following this analysis, I will use the same count data as response, and use the same 4 categorical predictors which result in the 5 dummy column in X matrix (excluding intercept) in Slice Sampling. The simulated parameters from the posterior distribution and the summary are given in Figure 16 and Table 14 as follows.

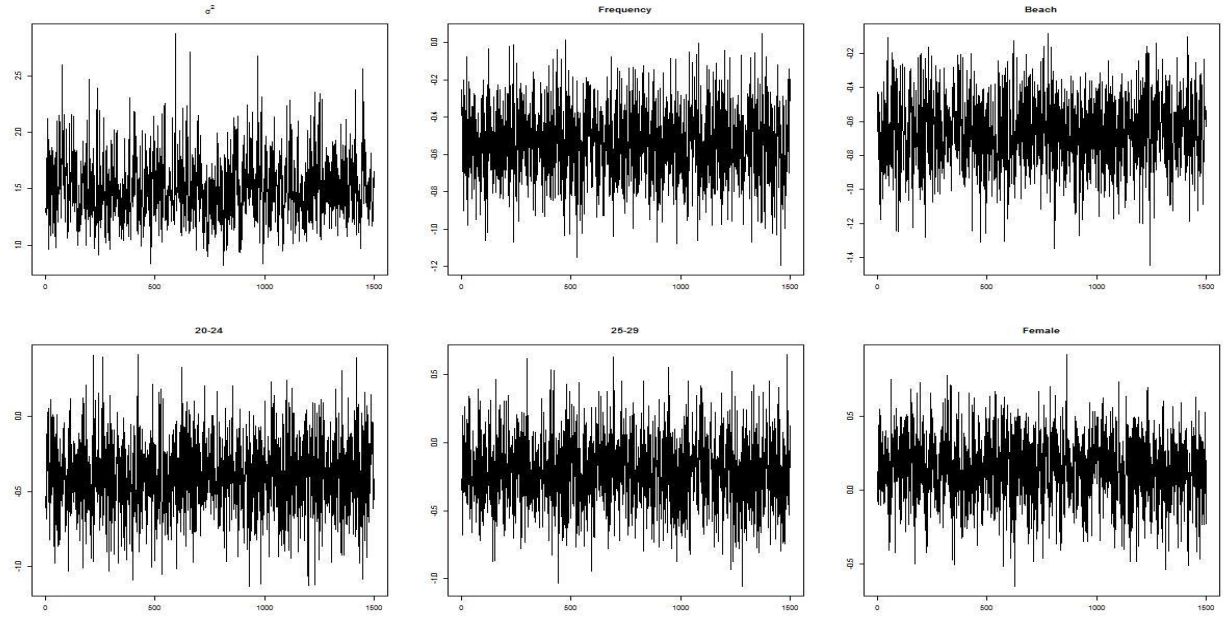

Figure 16. Parameters simulation (slice sampler) in Poisson regression

Parameter	Mean	95% Credible Set
σ^2	1.50149	(0, 2.012371)
Frequent	-0.5411584	(-0.9433834, -0.1500580)
Beach	-0.6652607	(-1.0873000, -0.2512824)
20-24	-0.4061987	(-0.89930632, 0.08860746)
25-29	-0.2082604	(-0.7215410, 0.3207603)
Female	0.1414067	(-0.336434, 0.574386)

Table 14 Posterior summary (slice sampler) in Poisson regression

3.6.5 Tree Ring Dataset

The tree-ring dataset (Originator: Swetnam, T.W., Caprio, A.C. and Lynch, A.M., https://www.ncdc.noaa.gov/paleo/study/5083) contains annual measurement between 837 AD and 1989 AD at Italian Canyon, New Mexico from PIFL Limber Pine at an altitude of 2894 m. There are two columns in the dataset, of which the first column is year and the second column is the ring data. Before fitting the stationary AR(1) model, I verify its stationarity using the test developed in Priestley and Rao (1969). Then, I divide the data into two parts. First part is training dataset, and the other test dataset have the last 30 observations which I will predict. The model is as follows, and the prior distributions, posterior distributions, and simulation procedures are exactly same with what I have illustrated in the Time Series chapter 3.5.

$$x_t - \mu = \phi(x_{t-1} - \mu) + \varepsilon_t; \varepsilon_t \sim N(0, \sigma^2)$$

The simulation results of the parameters and the posterior distribution summary are in Figure 17 and Table 15 respectively. The prediction result is shown in Figure 18.

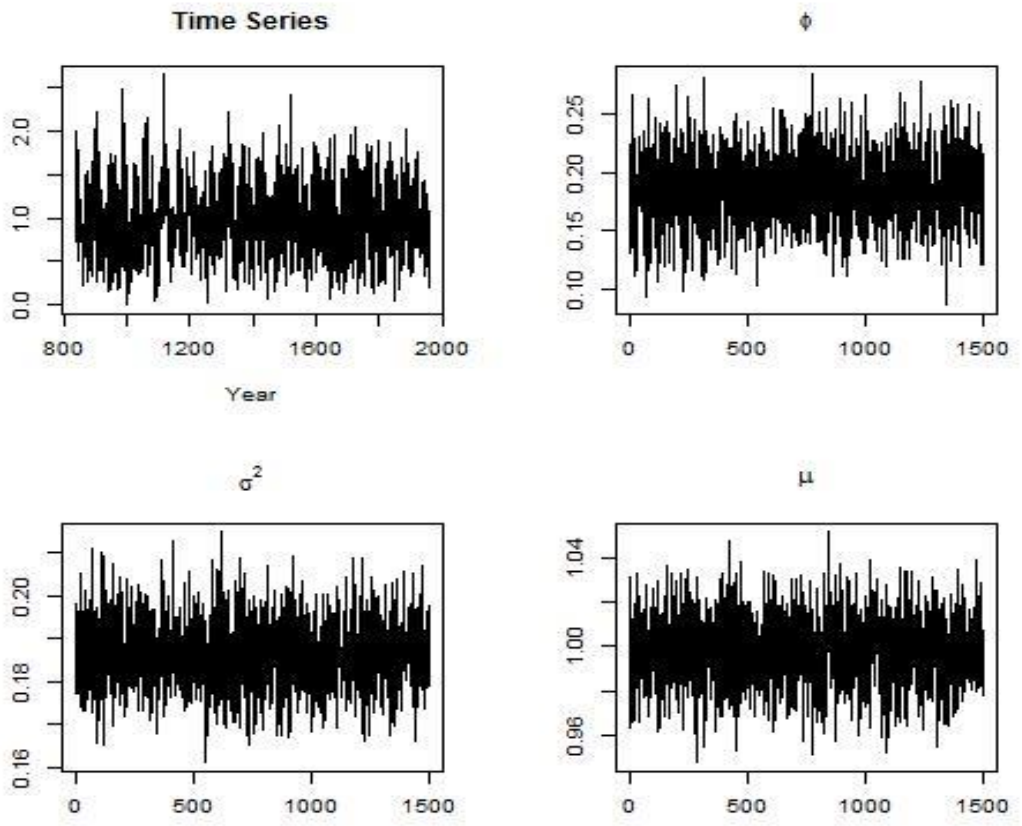

Figure 17. Parameters simulation in AR(1) time series

Summary	ϕ	σ^2	μ
Mean	0.1856	0.1865	0.9989
95% Credible Set	(0.1262,0.2435)	(0, 0.1996)	(0.9671,1.0299)

Table 15 Posterior summary in AR(1) time series

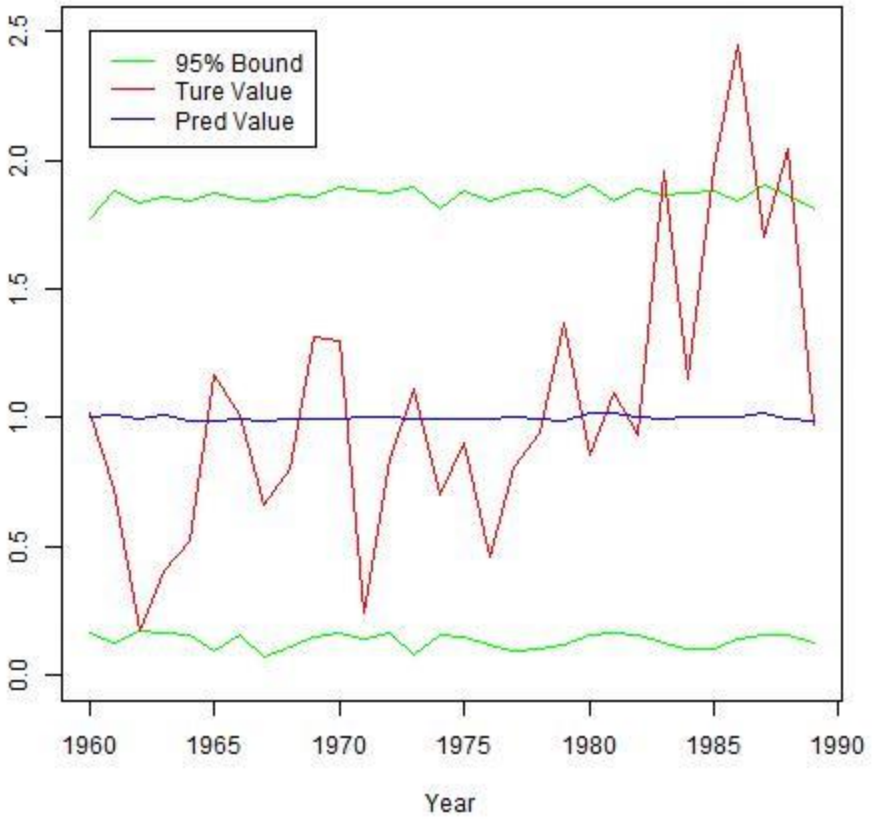

Figure 18. Prediction in AR(1) time series

Most of the true values lie within 95% predictive interval, which is good. But from the prediction result, we did not find the variance is increasing, and the reason is because that ϕ and σ^2 and both very small, close to 0. So the variance's increasing trend is not evident in this graph.

Chapter 4: Additional Topics in Bayesian Inference

4.1 Introduction

We have discussed the theoretical details underlying the Monte Carlo methods for Bayesian inference and gave many different examples of how to implement it. However, there are many other aspects of Bayesian modeling that one needs to be aware of before applying it to a problem and making decision based on the output. In the following, we discuss two of them.

4.2 Assessing Convergence in MCMC

As we have noted in Chapter 1, the validity of Monte Carlo method is dependent on our ability to draw a large number of independent samples from the target posterior so that empirical summaries converge to theoretical summaries. In case of MCMC, we need an additional convergence, because the samples we draw are not from the target distribution, but from a Markov chain that converges to the target distribution. Hence, it is important to check for convergence before we decide how many draws we are going to include in the MCMC. There are several different ways to assess convergence (Brooks and Roberts 1998, Plummer et al. 2006).

4.3 Model Comparison in Bayesian Inference

Consider a regression setting. One of the commonly encountered problem in regression is to decide on appropriate number of covariates. More covariates result in better fit but increases the dimension of parameter space and risks poor out-of-sample prediction. In likelihood-based methods, one uses criterion such as AIC or BIC (Akaike 1987; Burnham and Anderson 2004) that adds a penalty based on how many parameters are being used. No of parameters is not a

well-defined criterion in Bayesian method since, as we have seen with examples in Chapter 3, different sampling schemes may have different number of parameters for the same model, based on how we introduce auxiliary variables. There are alternative criteria in literature that are more suitable for a hierarchical model (Plummer 2008, Wilberg and Bence 2008).

Bibliography

Akaike, H., 1987. Factor analysis and AIC. Psychometrika, 52(3), pp.317-332.

Anderson, E., 1935. The irises of the Gaspe Peninsula. Bulletin of the American Iris society, 59, pp.2-5.

Azzalini, A. and Bowman, A.W., 1990. A look at some data on the Old Faithful geyser. Applied Statistics, pp.357-365.

Brooks, S.P. and Roberts, G.O., 1998. Convergence assessment techniques for Markov chain Monte Carlo. Statistics and Computing, 8(4), pp.319-335.

Burnham, K.P. and Anderson, D.R., 2004. Multimodel inference understanding AIC and BIC in model selection. Sociological methods & research, 33(2), pp.261-304.

Chib, S. and Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The American Statistician, 49(4), pp.327-335.

Gåsemyr, J., 2003. On an adaptive version of the Metropolis–Hastings algorithm with independent proposal distribution. Scandinavian Journal of Statistics, 30(1), pp.159-173.

Gelfand, A.E., 2000. Gibbs sampling. Journal of the American Statistical Association, 95(452), pp.1300-1304.

Gilks, W.R., 2005. Markov chain Monte Carlo. John Wiley & Sons, Ltd.

Haario, H., Saksman, E. and Tamminen, J., 1999. Adaptive proposal distribution for random walk Metropolis algorithm. Computational Statistics, 14(3), pp.375-396

Hand, D.J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E., 1994. A handbook of small data sets (Vol. 1). CRC Press.

Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. Applied logistic regression (Vol. 398). John Wiley & Sons.

Isaacson, D.L. and Madsen, R.W., 1976. Markov chains, theory and applications (Vol. 4). New York: Wiley.

Madsen, M., 1976. Statistical analysis of multiple contingency tables. Two examples. Scandinavian Journal of Statistics, pp.97-106.

Neal, R.M., 2003. Slice sampling. Annals of statistics, pp.705-741.

Plummer, M., Best, N., Cowles, K. and Vines, K., 2006. CODA: Convergence diagnosis and output analysis for MCMC. R news, 6(1), pp.7-11.

Plummer, M., 2008. Penalized loss functions for Bayesian model comparison. Biostatistics, 9(3), pp.523-539.

Priestley, M.B. and Rao, T.S., 1969. A test for non-stationarity of time-series. Journal of the Royal Statistical Society. Series B (Methodological), pp.140-149.

Shapiro, S.S. and Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), pp.591-611.

Weintraub, R., 1962. The birth rate and economic development: An empirical study. Econometrica: Journal of the Econometric Society, pp.812-817

Wilberg, M.J. and Bence, J.R., 2008. Performance of deviance information criterion model selection in statistical catch-at-age analysis. Fisheries Research, 93(1), pp.212-221.