Gibbs Sampler and Its Applications

Siqi Wang

May 1, 2015

Abstract

In this paper, we study the Gibbs sampler algorithm and explore some of its applications. First, we give the general outline of the algorithm and we find that a Markov Chain is constructed through the algorithm. Second, we explain that under certain conditions, the Markov Chain would have a unique stationary distribution. We also prove that the distribution satisfied the Detailed Balance condition is a stationary distribution of the Markov Chain. Later this results help us to make the proof that the target distribution is the only stationary distribution of the Markov Chain in Gibbs sampling. Finally we introduce a topic model called the LDA model and how can we fit the LDA model with Gibbs sampler. Several examples are given through the paper to help to get a better comprehension about this algorithm.

1 Introduction

The Gibbs Sampler (or Gibbs Sampling) is one of the commonly used Markov Chain Monte Carlo (MCMC) algorithms. MCMC algorithms, a kind of modified Monte Carlo method for simulation, are applied when sampling directly from the target distribution is not accessible. This technique burgeoned in the early 90's. A common feature of samples obtained using MCMC methods is that each sample value correlates with the nearby sample values. Thus the algorithm produces a sequence of sample values from a Markov Chain whose stationary distribution is the target distribution. Gibbs sampling is usually used to sample from posterior distributions in Bayesian analysis.

In Section 2, some basic definitions are presented and the Gibbs sampling algorithm scheme is described. We use an example to illustrate this algorithm then. Section 3 examines whether the Gibbs sampler converges to the target distribution. The section also discusses that Gibbs Sampling is well-suited to coping with incomplete information and an example under the situation of missing information is shown. Section 4 introduces an application of Gibbs sampler for statistical inference in topic modelling. An concrete example is presented as well.

2 Gibbs Sampling Algorithm

2.1 Key Notations And Definitions

First, in order to better explain the algorithm, we introduce the definition of the Markov Chain.

Definition 2.1. *Let S be a finite set, called the state space, where elements are called states. A sequence of random variables $(X^{(0)}, X^{(1)}, ..., X^{(T)})$ is*

called a Markov Chain if for all t such that $1 \leq t \leq T$*, the probability*

$$P(X^{(t)} = x_t | X^{(0)} = x_0, X^{(1)} = x_1, ..., X^{(t-1)} = x_{t-1}) = Pr(X^{(t)} = x_t | X^{(t-1)} = x_{t-1})$$

where $x_0, x_1, x_2, ..., x_t \in S$*. [?]*

From Definition 2.1 we know that in a Markov Chain, the current state is related and only related to the state of the last previous step. Since Gibbs sampling is a general method of probabilistic inference which usually dealing with the situation in the continuous space with joint random variables, we will now review some probability knowledge including the concept of distribution function and density function.

Definition 2.2. *Consider real-valued random variables* $X_1, X_2, ..., X_n$ *defined on the same sample space* Ω*. The joint probability distribution for* $X_1, X_2, ..., X_n$ *is described using a distribution function, defined as follows:*

$$F_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n) = P(X_1 \leqslant x_1, X_2 \leqslant x_2, ..., X_n \leqslant x_n).$$

A density function for the joint probability distribution, denoted as $f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n)$*, satisfies*

$$P(a_1 \leqslant X_1 \leqslant b_1, a_2 \leqslant X_2 \leqslant b_2, ..., a_n \leqslant X_n \leqslant b_n) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} ... \int_{a_n}^{b_n} f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n) dx_n ... dx_2 dx_1 \quad (1)$$

If such a density exists, we say $X_1, X_2, ..., X_n$ *are jointly distributed with joint density f. [?]*

An important distribution for estimating each parameter in the Gibbs sampler is the distribution of this parameter based on the information of other parameters, so now we need to introduce the idea of conditional distribution.

Definition 2.3. *Consider jointly distributed random variables $X_1, X_2, ..., X_n$ with distribution function $F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$ and the density function $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$. The conditional distribution function [?] of X_i ($i = 1, 2, ..., n$) given that $X_1 = x_1, ..., X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, ..., X_n = x_n$ is*

$$F_{X_i|X_1,...,X_{i-1},X_{i+1},...,X_n}(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = \int_{-\infty}^{x_i} \frac{f_{X_1,X_2,...,X_n}(x_1, ..., x_{i-1}, t, x_{i+1}, ..., x_n)}{f_{X_1,...,X_{i-1},Xi+1,...,X_n}(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)} dt \quad (2)$$

Now we define the "full conditional distribution" which we often used later in this paper.

Definition 2.4. *Consider jointly distributed random variables $X_1, X_2, ..., X_n$ defined on the same sample space Ω with distribution function $F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$. According to equation (2) in Definition 2.3, we define the conditional density function of X_i ($i = 1, 2, ..., n$) given $X_1 = x_1, ..., X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, ..., X_n = x_n$ as*

$$f_{X_i|X_1,...,X_{i-1},X_{i+1},...,X_n}(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = \frac{f_{X_1,X_2,...,X_n}(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n)}{f_{X_1,...,X_{i-1},Xi+1,...,X_n}(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)} \quad (3)$$

The corresponding conditional distribution of X_i that depends on all the other random variables is called the full conditional distribution. [?]

2.2 The Algorithm

Suppose the random vector $(X_1, X_2, ..., X_n)$ has a joint density $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$. Suppose also that we can simulate from the corresponding full conditional densities $f_{X_i|X_i^-}(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$ where $X_i^- = X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$ for all $i = 1, 2, ..., n$. The multistage Gibbs sampling scheme is:

1. Select initial values $X^{(0)} = (X_1^{(0)}, X_2^{(0)}, ..., X_n^{(0)}) = (x_1^{(0)}, x_2^{(0)}, ..., x_n^{(0)})$.

2. Set the index $t = 0$.

3. Generate random draws $x_1^{(t+1)}, x_2^{(t+1)}, ..., x_n^{(t+1)}$ where

   - sample $x_1^{(t+1)}$ from the full conditional distribution whose density is $f_{X_1^{(t+1)}}(x_1) = f_{X_1|X_1^-}(x_1|x_2^{(t)}, ..., x_n^{(t)})$
   - sample $x_2^{(t+1)}$ from the full conditional distribution whose density is $f_{X_2^{(t+1)}}(x_2) = f_{X_2|X_2^-}(x_2|x_1^{(t+1)}, x_3^{(t)}, ..., x_n^{(t)})$
   - sample $x_3^{(t+1)}$ from the full conditional distribution whose density is
     $f_{X_3^{(t+1)}}(x_3) = f_{X_3|X_3^-}(x_3|x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t)}, ..., x_n^{(t)})$
     ...
   - sample $x_n^{(t+1)}$ from the full conditional distribution whose density is
     $f_{X_n^{(t+1)}}(x_n) = f_{X_n|X_n^-}(x_n|x_1^{(t+1)}, x_2^{(t+1)}, ..., x_{n-1}^{(t+1)})$

4. Set $x^{(t+1)} = (x_1^{(t+1)}, ..., x_n^{(t+1)})$

5. Set $t = t + 1$ and repeat Step 3.

The algorithm is designed according to the work of Casella and George. [**?**]

Thus the Gibbs sampler generates a Markov Chain $(X^{(0)}, X^{(1)}, ..., X^{(T)})$ by construction. To better visualize what does the algorithm does, we provide an example made by Gelman. [**?**]

Example 1. Consider a jointly distributed random vector (X_1, X_2) with a bivariate normal distribution $N(\mu, \Sigma)$ as the target distribution where $\mu =$

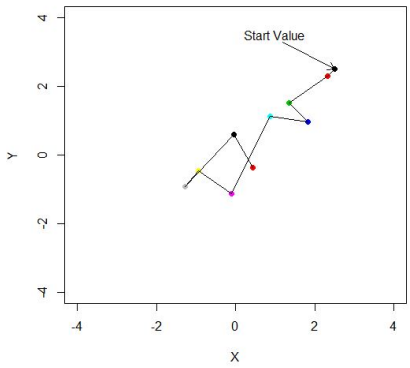


Figure 1: First 10 iterations start from (2.5,2.5)

$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Then the joint density function is

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \mathrm{e}^{\frac{-1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)}$$

. Thus we can get the marginal densities

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_2 = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x_1^2/2}$$

,

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1 = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x_2^2/2}$$

. The conditional densities are

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \mathrm{e}^{\frac{-(x_1-\rho x_2)^2}{2(1-\rho^2)}}$$

,

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} = \frac{1}{\sqrt{2\pi(1-\rho^2)}} \mathrm{e}^{\frac{-(x_2-\rho x_1)^2}{2(1-\rho^2)}}$$

. Now we show how to use these conditional distributions to sample from the target distribution by using the Gibbs sampler (the R implementation

can be found in Appendix A):

1. Take the initial value $(X_1^{(0)}, X_2^{(0)}) = (x_1^{(0)}, x_2^{(0)})$, also choose the iteration time T.

2. Set the index $t = 0$.

3. Generate random draws $x_1^{(t+1)}, x_2^{(t+1)}$ from $f_{X_1|X_1^-}(x_1|x_2^{(t)})$ and $f_{X_2|X_2^-}(x_2|x_1^{(t+1)})$ where

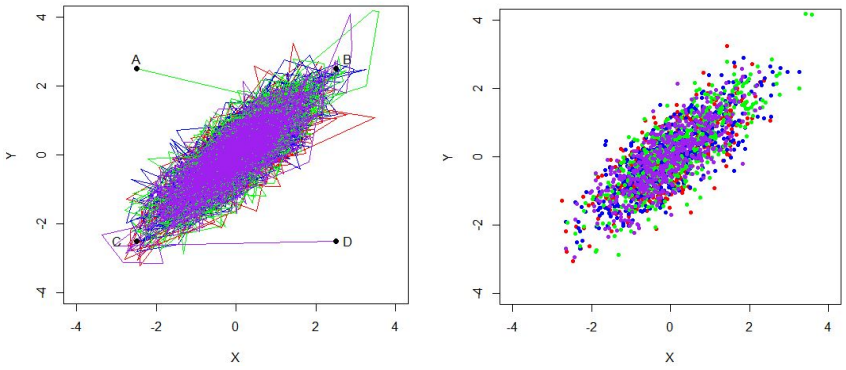
   - $f_{X_1|X_1^-}(x_1|x_2^{(t)})$ is $N(\rho x_2^{(t)}, 1-\rho^2)$,
   - $f_{X_2|X_2^-}(x_2|x_1^{(t+1)})$ is $N(\rho x_1^{(t+1)}, 1-\rho^2)$.

4. Set $t = t + 1$ and repeat Step 3 unless $t = T$.

Figure 2a and Figure 2b show samples by Gibbs sampling with $\rho = 0.8$. In Figure 1, all 1000 iterations (which means $T = 1000$) are shown and in Figure 2, the last 500 iterations are presented. We discard the first 500 iterations to make sure that all the rest are good approximates of the target distribution. Figure 3 shows the 1000 samples from the target distribution which is the bivariate normal distribution with $\rho = 0.8$ (we make the random samples in r since we can find the bivariate normal distribution). Comparing Figure 2(b) and Figure 3, we can clearly see that Gibbs sampler is a great approximation about the target distribution.

3 Stationary Distributions of Markov Chains in Gibbs Sampling

Although Gibbs sampling builds a Markov Chain through running the algorithm, it has not yet been shown that we can use this algorithm to sample from the target distribution. In this section the stationary distribution of Markov Chains is introduced and we show that the target distribution is the



(a) (b)

Figure 2: (a):1000 iterations start from 4 starting points: A, B, C, D; (b) Last 500 samples in each of 4 runs

unique stationary distribution of the Markov Chain defined by the Gibbs sampler. In other words, the distribution obtained from Gibbs sampling converges to the target distribution as the iteration number $T \to \infty$.

3.1 Some Definitions and Terminology

To make our proof more understandable, we mainly talk about discrete-state Markov chains where the state space in finite.

Definition 3.1. *A Markov chain is defined by its transition probabilities $P(x, y)$ (or we can call it as transition kernel) where x, y are elements in the state space S. At any step t,*

$$P(x, y) = P(X^{(t)} = y | X^{(t-1)} = x). \tag{4}$$

Suppose that the state space S has n elements ($n \in \mathbb{N}$) and these elements are ordered from 1 to n. We can construct a $n \times n$ transition matrix P. Suppose we order all the elements in S and $x \in S$ is the i^{th} element and $y \in S$ is the

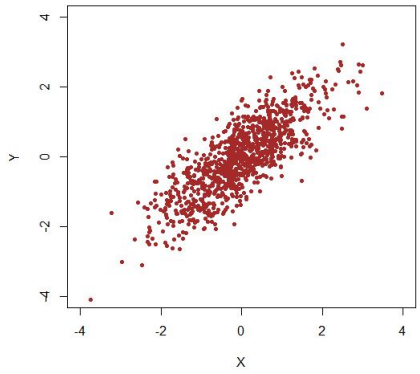


Figure 3: 1000 random draws from the target distribution

j^{th} *element. Then, the* $(i,j)^{\text{th}}$ *entry* P_{ij} *of* P *is*

$$P_{ij} = P(x,y) \qquad (i,j \leqslant n).$$

[?]

A *time-independent Markov chain* means the transition matrix P does not change at any step t. In contrast, a *time-dependent Markov Chain* does not have a consistent transition matrix over time; thus, at step t, we denote the transition matrix as P_t. We mainly use the time-independent Markov Chain to simplify explanations. However the definition for time-dependent Markov Chain is very similar. Just think of a time-independent Markov Chain as a special case of time-dependent Markov Chain where transition matrices P_t of any step t are all equal to P.

Let $\pi_t(x) = P(X^{(t)} = x)$ denote the probability that the chain is in state x at step t and let π_t be the row vector which describes state probabilities at time t. The starting probability vector (the probability mass function of the

initial distribution) is denoted π_0. Then, every step t,

$$\sum_{x \in S} \pi_t(x) = 1$$

.

We can now calculate the probability that the chain is in state y at step $t+1$:

$$\begin{aligned} \pi_{t+1}(y) &= P(X^{(t+1)} = y) \\ &= \sum_{x \in S} P(X^{(t+1)} = y | X^{(t)} = x) \cdot P(X^{(t)} = x) \\ &= \sum_{x \in S} P(x, y) \cdot \pi_t(x) \end{aligned}$$

Based on Definition 3.1, we can write the equation above in the matrix form

$$\pi_{t+1} = \pi_t P, \tag{5a}$$

and we can prove that

$$\pi_t = \pi_0 P^t. \tag{5b}$$

Proof. Given the initial vector π_0, when $t = 1$, we have

$$\pi_1 = \pi_0 P.$$

The claim is satisfied.

Suppose the claim is true for step $t-1$ ($t > 1$), then at step t, we find

$$\begin{aligned}\pi_t &= \pi_{t-1}P \\ &= (\pi_0 P^{t-1})P \\ &= \pi_0 P^t.\end{aligned}$$

Thus, Equation 5b is proved by induction. □

Then P^n can be seen as a n^{th} step transition matrix and according to equation 5b, it is clear that the $(i,j)^{\text{th}}$ entry $(P^n)_{ij}$ of P^n is

$$(P^n)_{ij} = P(X^{(t+n)} = y | X^{(t)} = x) \quad \forall t \in \mathbb{N}.$$

3.2 Stationary Distribution

An important property of the Markov Chain is the stationary distribution. We now introduce its definition.

Definition 3.2. *Let π^* be a probability mass function (written as a row vector), and let P be the transition matrix of a Markov chain. The probability mass function π^* is the stationary distribution of the chain if*

$$\pi^* = \pi^* P. \tag{6}$$

[?]

Note that

$$\pi^* = \pi^* P^t \quad \forall t \in \mathbb{N}$$

We need to have attention that not all kinds of Markov Chains have stationary distributions. In other words, only those Markov Chains that

satisfy certain criteria equip stationary distributions. To explain it, we first need to know what irreducibility and aperiodicity are.

Definition 3.3. *Given a Markov Chain with transition matrix P, if there exists a positive integer n_{ij} for each ij^{th} entry such that $(P^{n_{ij}})_{ij} > 0$, then the chain is irreducible. If for any state x in the state space S, $gcd\{n \in \mathbb{N}^+ : P^n(x,x) > 0\} = 1$, then we say the Markov chain is aperiodic.[?]*

Theorem 3.4. *A Markov Chain with transition matrix P has an unique stationary distribution π^* if and only if the chain is irreducible and aperiodic. That for a irreducible and aperiodic chain, it means that given an arbitrary starting probability vector π_0, we can get*

$$\pi_0 P^n \to \pi^* \qquad as\ n \to \infty \tag{7}$$

Proof. See page 1710-1715 of Tierney [?] for a proof of Theorem 3.4. □

Theorem 3.4 can help us to determine whether the Markov chain had an unique stationary distribution. At the same time, when there is a probability distribution, we also want to determine whether it is a stationary distribution of our Markov Chain.

Lemma 3.5. *Consider a Markov Chain with transition matrix P and finite state space S. Let x, y be two arbitrary elements in S. Let π^* be a probability vector. Then, π^* is a stationary distribution of the Markov chain if*

$$P(x,y)\pi^*(x) = P(y,x)\pi^*(y) \tag{8}$$

We call this relationship the Detailed Balance Condition. [?]

Proof. Consider the vector $\pi^* P$. For an arbitrary state $y \in S$, by (8) we

have

$$\begin{aligned}(\pi^* P)(y) &= \sum_{x \in S} \pi^*(x) P(x, y) \\ &= \sum_{x \in S} \pi^*(y) P(y, x) \\ &= \pi^*(y) \sum_{x \in S} P(y, x) \\ &= \pi^*(y)\end{aligned}$$

In other words, $\pi^* = \pi^* P$. □

By Theorem 3.4 and Lemma 3.5, for any irreducible and aperiodic Markov chain, if a probability mass function π^* satisfies the detailed balance condition, then for all π_0,

$$\lim_{n \to \infty} \pi_0 P^n = \pi^* \tag{9}$$

The detailed balanced in the continuous space is analogous to (8) which we define in the discrete space. In the continuous case, we would have a transition kernel $P(x, y)$ that satisfies

$$\int P(x, y) dy = 1$$

where $P(x, y) = f_{X^{(t)}|X^{(t-1)}}(y|x)$ where $f_{X^{(t)}|X^{(t-1)}}$ is the conditional density function and x describes the state information of step $t - 1$ and y describes the state information of step t.

The continuous extension of the formula in (5a) is

$$\pi_t(y) = \int \pi^*_{t-1}(x) P(x, y) dx. \tag{10}$$

The formula in (6) for the stationary distribution π^* will be modified to

$$\pi^*(y) = \int \pi^*(x)P(x,y)dx. \tag{11}$$

In Tierney's proof [8, p.1710-1715], we can see that Theorem 3.4 and Lemma 3.5 still hold for the continuous extension.

3.3 Stationary Distribution in the Gibbs Sampler

Let's consider a random vector $X = (X_1, X_2, ..., X_n)$ with joint distribution π. To simplify the scenario, let's assume X has a finite state space S. We know the corresponding full conditional probabilities $P_{X_i|X_i^-}(x_i|x_i^-)$ for all $i = 1, 2, ..., n$. Suppose π is unknown and we want to use the Gibbs sampler to approximate π. In other word, π is our target distribution.

Lemma 3.6. *Consider the Markov Chain defined in the Gibbs sampling algorithm with state space S. The target distribution of the algorithm satisfies the detailed balance condition in this Markov Chain. [?]*

Proof. To prove the lemma, it is sufficient to show that the transition between two arbitrary states $x, y \in S$ at step t satisfies the detailed balance condition. Assume $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ are two possible states in the state space S. Then at an arbitrary step t, X is at the state x. By the construction of the Gibbs sampler, each step contains n substeps and these n substeps form a time-dependent Markov Chain. Thus the transition matrix $P(x, y)$ can be factorized into n transition matrices

$$P(x,y) = P_1(x, y^{(1)})P_2(y^{(1)}, y^{(2)})...P_n(y^{(n-1)}, y) \tag{12}$$

where $y^{(i)} = (y_1, y_2, ..., y_i, x_{(i+1)}, ..., x_{(n)})$, and $P_i(y^{(i-1)}, y^{(i)}) = P_{X_i|X_i^-}(y_i|y_1, ..., y_{i-1}, x_{i+1}, ..., x_n)$ for $i = 1, 2, ..., n-1$.

Note that for any states x, y, we have $P_i(x, y) = 0$ if $(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) \neq (y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$.

To prove the detailed balance for time-dependent Markov chain, it is sufficient to show detailed balance for each transition matrix. For an arbitrary transition matrix P_i, if states x, y satisfy that $(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = (y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$, then

$$\begin{aligned}
\pi(x)P_i(x,y) &= P_X(x_1, ..., x_n)P_{X_i|X_i^-}(y_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) \\
&= P_X(x_1, ..., x_n)\frac{P_X(x_1, ...x_{i-1}, y_i, x_{i+1}, ..., x_n)}{P_{X_i^-}(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)} \\
&= P_{X_i|X_i^-}(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)P_X(x_1, ...x_{i-1}, y_i, x_{i+1}, ..., x_n) \\
&= P_{X_i|X_i^-}(x_i|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)P_X(y_1, ...y_{i-1}, y_i, y_{i+1}, ..., y_n) \\
&= P_i(y,x)\pi(y),
\end{aligned}$$

while for states x, y such that $(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) \neq (y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$, the detailed balance is trivially fulfilled.

Since P_i is arbitrarily chosen, each of $P_1, P_2, ..., P_n$ satisfies the detailed balance condition.

Thus the detailed balance condition is satisfied.

□

The proof is made with the hint given by Besag's work. [**?**]

The proof above is for the high-dimensional case which seems very complicated. To help understand Lemma 3.6, now we consider a bivariate random variable $X = (X_1, X_2)$ with state space S and $x = (x_1, x_2)$, $y = (y_1, y_2)$ are two possible states in S. We will reprove Lemma 3.6 in this case

Proof. The transition matrix $P(x, y)$ can be factorized into 2 transition matrices

$$P(x,y) = P_1(x,y')P_2(y',y) \tag{13}$$

where $y' = (y_1, x_2)$ is the result of the intermediate step and $P_1(x, y') = P_{X_1|X_2}(y_1|x_2)$, $P_2(y', y) = P_{X_2|X_1}(y_2|y_1)$

Note that for any states x, y, we have $P_1(x, y) = 0$ if $x_2 \neq y_2$, $P_2(x, y) = 0$ if $x_1 \neq y_1$.

Now we need to show that detailed balance holds for each transition matrix. For P_1, if states x, y satisfy that $x_2 = y_2$, then

$$\begin{aligned}
\pi(x)P_1(x, y) &= P_X(x_1, x_2)P_{X_1|X_2}(y_1|x_2) \\
&= P_X(x_1, x_2)\frac{P_X(y_1, x_2)}{P_{X_2}(x_2)} \\
&= P_{X_1|X_2}(x_1|x_2)P_X(y_1, x_2) \\
&= P_{X_1|X_2}(x_1|y_2)P_X(y_1, y_2) \\
&= P_1(y, x)\pi(y),
\end{aligned}$$

while for states x, y such that $x_2 \neq y_2$, the detailed balance is trivially fulfilled.

Similarly for P_2, if states x, y satisfy that $x_1 = y_1$, then

$$\begin{aligned}
\pi(x)P_2(x, y) &= P_X(x_1, x_2)P_{X_2|X_1}(y_2|x_1) \\
&= P_X(x_1, x_2)\frac{P_X(x_1, y_2)}{P_{X_1}(x_1)} \\
&= P_{X_2|X_1}(x_2|x_1)P_X(x_1, y_2) \\
&= P_{X_2|X_1}(x_2|y_1)P_X(y_1, y_2) \\
&= P_2(y, x)\pi(y),
\end{aligned}$$

while for states x, y such that $x_1 \neq y_1$, the detailed balance is trivially fulfilled.

Thus the detailed balance condition is satisfied.

□

According to Theorem 3.4, we know that Gibbs sampling would finally converge to the target distribution as long as the chain is aperiodic and irreducible. However we also worry about the situation such that a "bad" starting point misleads the direction of the following iterations. Now let's discover whether the initial value would influence the convergence of Markov chain constructed by the Gibbs sampler.

Corollary 3.7. *The Markov Chain built through the Gibbs sampling converges to the target distribution with any start point* π_0.

Proof. Casella and George [4, p.169-170] point out that the Markov Chain defined by the Gibbs sampling algorithm is irreducible and aperiodic. Then by Theorem 3.4, we have that for any initial distribution π_0, there exists an unique distribution π^* such that

$$\lim_{n\to\infty} \pi_0 P^n = \pi^*$$

where P is the transition matrix of the Markov chain

In Lemma 3.6, we have proved that the target distribution satisfies the detailed balance condition. Thus the unique stationary distribution is the target distribution. □

By Theorem 3.4, we know that when the Markov chain built through Gibbs sampling is irreducible and aperiodic, it would converge to a unique stationary distribution, which is the target distribution we want to sample from. At the same time, we do not need to worry about where do we start the algorithm. Thus the Gibbs sampler is a valid tool for sampling from the target distribution.

Though theoretically we need to know the full conditional distribution for the sampling, in fact we can drop the normalizing constant in its density when we use the algorithm. The normalizing constant is the constant that

scales the the density so that its integral is 1. For instance, in Example 1, the normalizing constant of $f_{X_1|X_2}(x_1|x_2)$ is $\frac{1}{\sqrt{2\pi(1-\rho^2)}}$. In other words, we may only need to know a function proportional to the full conditional density for the sampling process. The next example shows this property.

Example 2. The 19^{th} century German physician Carl Reinhold August Wunderlich once stated that the mean healthy human body temperature is 37 °C(98.6 °F). In 1992, Mackowiak et al. collected 130 healthy people's data [**?**]: their body temperature, gender and heart rate. To use these data to check Wunderlich's statement, Gibbs sampling can be applied to simulate body temperature. If the mean of the posterior distribution is close to 98.6 °F then we accept Wunderlich's claim. To conduct the hypothesis test, we need the variance of the posterior distribution.
Suppose the i^{th} person's body temperature is y_i $(i = 1, ..., n)$ and we assume that each y_i is normally distributed as

$$y_i \sim N(\mu, \sigma^2)$$

The prior distribution of μ and σ^2 are

$$\mu \sim N(\mu_0, \sigma_0^2) \quad \sigma^2 \sim IG(a_0, b_0) \quad (\mu_0, \sigma_0^2 \in \mathbb{R}, a_0, b_0 > 0)$$

$IG(a_0, b_0)$ stands for the inverse-gamma distribution with parameter a_0, b_0 such that the density function is $f(\sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-a_0-1}\mathrm{e}^{b_0/\sigma^2}$.
Define $y = (y_1, ..., y_n)$. We can find that

$$\begin{aligned} f(\mu, \sigma^2|y) &= \frac{f(y|\mu, \sigma^2)f(\mu)f(\sigma^2)}{\int_\mu \int_{\sigma^2} f(y|\mu, \sigma^2)f(\mu)f(\sigma^2)d\sigma^2 d\mu} \\ &\propto f(y|\mu, \sigma^2)f(\mu)f(\sigma^2) \end{aligned}$$

Since it is not possible to compute the integral $\int_\mu \int_{\sigma^2} f(y|\mu, \sigma^2)f(\mu)f(\sigma^2)d\sigma^2 d\mu$, we cannot find the posterior distribution. Thus we decide to use the Gibbs

sampler as an alternative way. To apply the Gibbs sampler, we need to find the full conditionals $f(\mu|\sigma^2, y)$,$f(\sigma^2|\mu, y)$. It is easy to show that

$$\begin{aligned}
f(\mu|\sigma^2, y) &\propto f(y|\mu, \sigma^2)f(\mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}} \mathrm{e}^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}} \\
&\propto \mathrm{e}^{-\sum_{i=1}^{n} \frac{(y_i-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}} \\
&= \mathrm{e}^{-\frac{1}{2}(\sum_{i=1}^{n} \frac{(y_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2})} \\
&= \mathrm{e}^{-\frac{1}{2\sigma^2\sigma_0^2}(\sigma_0^2 \sum_{i=1}^{n} (y_i-\mu)^2 + \sigma^2(\mu-\mu_0)^2)} \\
&= \mathrm{e}^{-\frac{1}{2\sigma^2\sigma_0^2}(\sigma_0^2 \sum_{i=1}^{n} (y_i^2-2y_i\mu+\mu^2) + \sigma^2(\mu^2-2\mu\mu_0+\mu_0^2))}
\end{aligned}$$

Now we multiply the $2y_i\mu$ term in the summation by $\frac{n}{n}$ in order to get the equation in terms of $\overline{y}$ where $\overline{y}$ is the data mean:

$$\begin{aligned}
f(\mu|\sigma^2, y) &\propto \mathrm{e}^{-\frac{1}{2\sigma^2\sigma_0^2}(\sigma_0^2 \sum_{i=1}^{n} (y_i^2-2\frac{n}{n}y_i\mu+\mu^2) + \sigma^2(\mu^2-2\mu\mu_0+\mu_0^2))} \\
&= \mathrm{e}^{-\frac{1}{2\sigma^2\sigma_0^2}(\sigma_0^2 \sum_{i=1}^{n} y_i^2 - \sigma_0^2 2n\overline{y}\mu + \sigma_0^2 n\mu^2 + \mu^2\sigma^2 - 2\mu\mu_0\sigma^2 + \mu_0^2\sigma^2)}.
\end{aligned}$$

We can represent terms that do not contain μ by a constant. Let $k = \mu_0^2\sigma^2 + \sigma_0^2 \sum_{i=1}^{n} y_i^2$; then,

$$\begin{aligned}
f(\mu|\sigma^2, y) &\propto \mathrm{e}^{-\frac{1}{2\sigma^2\sigma_0^2}(\mu^2(\sigma^2+\sigma_0^2 n) - 2\mu(\mu_0\sigma^2+\sigma_0^2 n\overline{y}) + k)} \\
&= \mathrm{e}^{-\frac{1}{2}(\mu^2(\frac{\sigma^2+\sigma_0^2 n}{\sigma^2\sigma_0^2}) - 2\mu(\frac{\mu_0\sigma^2+\sigma_0^2 n\overline{y}}{\sigma^2\sigma_0^2}) + k)}.
\end{aligned}$$

Now we multiply the exponent by $\frac{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}$,

$$\begin{aligned} f(\mu|\sigma^2, y) &\propto \mathrm{e}^{-\frac{1}{2}(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2})(\mu^2-2\mu(\frac{\frac{\mu_0}{\sigma_0^2}+\frac{n\overline{y}}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}})+k)} \\ &= \mathrm{e}^{-\frac{1}{2}(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2})(\mu-(\frac{\frac{\mu_0}{\sigma_0^2}+\frac{n\overline{y}}{\sigma^2}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}))^2} \\ &\propto \frac{1}{\sqrt{2\pi\omega\frac{\sigma^2}{n}}}\mathrm{e}^{-\frac{(\mu-(\omega\overline{y}+(1-\omega)\mu_0))^2}{2\omega\frac{\sigma^2}{n}}} \end{aligned}$$

where $\omega = \frac{\sigma_0^2}{\frac{\sigma^2}{n}+\sigma_0^2}$.
Thus,

$$\mu|\sigma^2, y \sim N(\omega\overline{y} + (1-\omega)\mu_0, \omega\frac{\sigma^2}{n})$$

We find that the mean of this distribution is a weighted average of the prior mean μ_0 and the data mean $\overline{y}$.

Similarly, we can find that

$$\begin{aligned}
f(\sigma^2|\mu, y) &\propto f(y|\mu, \sigma^2) f(\sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \frac{b_0^{a_0}}{\Gamma(a_0)} \sigma^{-2(a_0+1)} \mathrm{e}^{\frac{b_0}{\sigma^2}} \\
&\propto \prod_{i=1}^{n} \frac{1}{\sigma} \mathrm{e}^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \sigma^{-2(a_0+1)} \mathrm{e}^{\frac{b_0}{\sigma^2}} \\
&= \frac{1}{\sigma^2} \mathrm{e}^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \sigma^{-2(a_0+1)} \mathrm{e}^{\frac{b_0}{\sigma^2}} \\
&= \sigma^{-2(a_0+\frac{n}{2}+1)} \mathrm{e}^{-(\frac{b_0}{\sigma^2} + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2})} \\
&= \sigma^{-2(a_0+\frac{n}{2}+1)} \mathrm{e}^{-\frac{2b_0 + 2\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}}{2\sigma^2}} \\
&= \sigma^{-2(a_0+\frac{n}{2}+1)} \mathrm{e}^{-\frac{b_0 + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}}{\sigma^2}} \\
&\propto \frac{b_1^{a_1}}{\Gamma(a_1)} \sigma^{-2a_1+1} \mathrm{e}^{-\frac{b_1}{\sigma^2}}
\end{aligned}$$

where $a_1 = a_0 + \frac{n}{2}, \quad b_1 = b_0 + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}$.
Thus,

$$\sigma^2|\mu, y \sim IG(a_0 + \frac{n}{2}, b_0 + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}).$$

Now we can develop the sampling scheme. Given Mackowiak's data of 130 healthy people's body temperature, we know that $n = 130$ in this case.

1. Initialize μ_0, σ_0^2 and determine the iteration number T and fixed values a_0 and b_0.

2. Set the index $t = 1$.

3. Calculate $\omega = \frac{\sigma_0^2}{\frac{\sigma_{t-1}^2}{n} + \sigma_0^2}$, $m = \omega\overline{y} + (1-\omega)\mu_0$, $s^2 = \omega\frac{\sigma^2}{n}$.

4. Generate μ_t from the distribution $N(m, s^2)$.

5. Calculate $a = a_0 + \frac{n}{2}$, $b = b_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu_t)^2$.

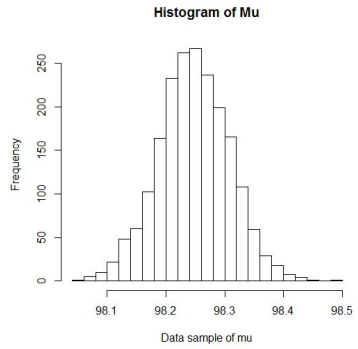
6. Generate τ from $Gamma(a, b)$.

7. Let $\sigma_t^2 = \frac{1}{\tau}$.

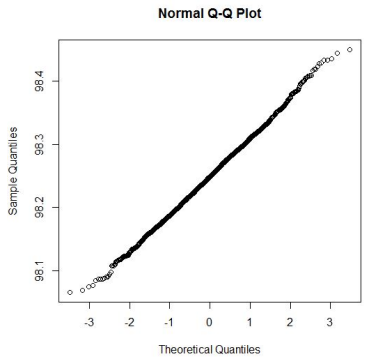
8. Set $t = t + 1$ and return to step 3 unless $t = T$.

After running the code for 5000 iterations with initial values $\mu_0 = 98.6$, $\sigma_0 = 0.5$, $a_0 = b_0 = 0.001$, we find that the mean of the sample data is approximately 98.25 °F which is about 36.8 °C. This result gives us a preliminary judgement that the body temperature in Wunderlich's statement is not accurate enough. More test can be made in the following process to check this judgement.

Figure 4(a) is the histogram of μ data from Gibbs sampling where we choose the latest value of μ after 5000 iterations and we obtain 2000 values of this kind. The reason for obtaining data this way (let's call it "the new method") instead of drawing the data with the last 2000 or 1000 values from the 5000 iterations is by the nature of a Markov Chain, consecutive data values are correlated. However, we need independent samples to estimate the distribution of μ. The new method is a common technique to obtain independent samples using Gibbs sampling. [**?**] We also estimated the mean and the variance of the distribution of μ. The mean is about 98.25 and the standard deviation is approximately 0.06. The normal quantile plot in Figure 4(b) indicates that μ is normally distributed. The mean body temperature in Wunderlich's statement is more than 3 standard deviations away from the posterior mean (posterior mean is the mean of the posterior distribution). Thus, we cannot accept the original statement. Mackowiak et al. has made the same conclusion and he did several tests to check this result in his research.[**?**] The R implementation is in Appendix B.



(a) Histogram of sample data drawn from the posterior distribution of μ



(b) Normal Quantile plot of the sample data

4 Application in Topic Modelling

Topic modelling is a subfield of machine learning which can be applied to many domains such as bioinformatics, political science and computational linguistics. In this paper, I introduce a specific model, Latent Dirichlet Allocation (LDA). I show how to fit LDA to documents using the Gibbs sampling algorithm.

4.1 Latent Dirichlet Allocation

LDA is a model introduced by Blei, Ng and Jordan [**?**] in 2003. It is for collections of grouped discrete data. A "collection" might be a collection of documents. The data values in a group are the words in a document. LDA is a three-level hierarchical Bayesian model[**?**], in which each group of the collection is described by a mixture over a finite set of latent topics, where each topic is a discrete distribution over the vocabulary of the collection. While LDA can be applied to any corpus of discrete data, in this paper, we think of a corpus as a collection of documents and the basic unit of discrete data is a word.

We introduce some notation and terminology before explaining the LDA model:

- A word is defined to be an element from the collection's vocabulary indexed by $\{1, ..., V\}$. A random word v from the corpus belongs to $\{1, ..., V\}$.
- The latent topics of the corpus are indexed by $\{1, ..., K\}$. We assume that for any corpus we are interested in, $K \geq 2$
- A corpus is a collection of D documents indexed by $\{1, 2, ..., D\}$.
- Document d $(1 \leq d \leq D)$ consists of N_d words. The corpus has $\sum_{d=1}^{D} N_d = N$ words in total. To sum up the corpus consists of words $w = (w_1, w_2, ..., w_N)$, where w_i is in some document d (for all $i \neq j$, w_i and w_j do not have to be different).

In Latent Dirichlet Allocation, a corpus (or we can call it a document collection) is generated by following procedure: Let corpus-level parameters $\alpha > 0$ and $\eta > 0$ be given. Let $K \geq 2$ be given,

1. For each topic $k \in \{1, ..., K\}$:

   (a) $\beta_k \sim Dir_V(\eta)$ where Dir_V denotes the V-dimensional dirichlet distribution

2. For each document $d \in \{1, ..., D\}$:

   (a) $\theta_d \sim Dir_K(\alpha)$

   (b) For each of the words w_i in d:

   - $P(z_i = k|\theta_d) = \theta_{d,k}$, so each z_i is chosen independently with parameter θ_d
   - $P(w_i = v|z_i = k, \beta_k) = \beta_{k,v}$

In the process above, β_k is the probability mass function over the fixed vocabulary that describes the topic k, which is dirichlet distributed in the prior. And θ_d is the probability mass function over topics in document d, z_i is the topic variable for word w_i in a certain document, α and η are hyperparameters of the symmetric Dirichlet distributions which are prior distributions for the parameters, θ_d $1 \leq d \leq D$ and β_k $1 \leq k \leq K$. Several assumptions are made in this model. First, we assume that each topic variable z's dimension is known and fixed (we have K topics in total). Second, β represents a $K \times V$ matrix such that $\beta_{k,v} = P(w = v|z = k)$ ($k \in \{1, ...K\}$, $v \in \{1, ...V\}$). β_k is the k^{th} row of β ($1 \leq k \leq K$).

Thus the random variable θ_d for each document d is K-dimensional and has the following probability density:

$$P(\theta_d|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha-1}, \tag{14}$$

where

$$\frac{1}{B(\alpha)} = \frac{\Gamma(\alpha)^K}{\Gamma(K\alpha)}.$$

Similarly, the parameter β_k for each topic k has a V-dimensional Dirichlet distribution:

$$P(\beta_k|\eta) = \frac{1}{B(\eta)} \prod_{v=1}^{V} \beta_{k,v}^{\eta-1}, \tag{15}$$

where

$$\frac{1}{B(\eta)} = \frac{\Gamma(\eta)^V}{\Gamma(V\eta)}.$$

Figure 5 gives a graphical representation of the generating a corpus using the LDA model.

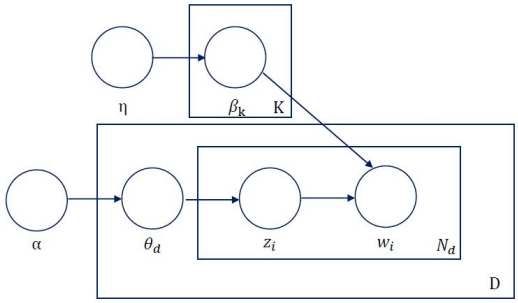


Figure 5: LDA Model by Graph

4.2 Inference by Gibbs Sampling

From Section 4.1 we can clearly see that $\{\beta_k\}_{1\leq k\leq K}$ and $\{\theta_d\}_{1\leq d\leq D}$ are parameters that have not been observed, hence which need to be estimated from the words in the document of the corpus. Several methods are introduced to fit LDA to a corpus such as EM (expectation-maximization) with expectation propagation[**?**], EM with variational inference [**?**] and Gibbs sampling. In this paper, we focus on how to estimate the parameters $\{\beta_k\}_{1\leq k\leq K}$ and $\{\theta_d\}_{1\leq d\leq D}$ by the Gibbs sampler.

Just like the problem we met in Example 2, it is not possible for us to figure out the posterior distribution exactly. The computation of the denominator in Equation (16) becomes our major computational problem. Thus, instead of representing $\{\beta_k\}$ and $\{\theta_d\}$ explicitly as parameters to be estimated, we use the advantage of Gibbs sampling and aim to evaluate the posterior distribution and then update the parameters through the algorithm. It requires us to reverse the defined generative process of LDA and obtain the posterior distributions of the latent variables in the model given the observed data.

$$p(z|w) = \frac{p(w,z)}{\sum_{z'} p(w,z')} \tag{16}$$

where $z = \{z_1, ..., z_N\}$.

Since each document has its own topic mixture and the method of updating parameters of each document is the same, here we will focus on evaluating the posterior distribution of one document, call it d.

According to the Bayes' Rule, we can express the posterior distribution as

$$P(z_i = k|z_{-i}, w) \propto P(w_i|z_i = k, z_{-i}, w_{-i})P(z_i = k|z_{-i}) \tag{17}$$

where z_{-i} is the assignment of all z_j such that $j \neq i$ and w_{-i} represents all words except w_i. On the right hand side of Equation 17, the first term is a likelihood function and the second term represents the prior.

For the first term on the right hand side, we can express it as

$$P(w_i|z_i = k, z_{-i}, w_{-i}) = \int P(w_i|z_i = k, \beta_k)P(\beta_k|z_{-i}, w_{-i})d\beta_k \tag{18}$$

In this equation, $P(w_i|z_i = k, \beta_k)$ is just β_{k,w_i} by definition (as we have explained in section 4.1). Then we can obtain the second term in the integral

of Equation 18 from the Bayes' rule

$$P(\beta_k|z_{-i}, w_{-i}) \propto P(w_{-i}|z_{-i}, \beta_k)P(\beta_k|z_{-i}) \tag{19}$$

Since $P(\beta_k)$ is $Dir_V(\eta)$ which is conjugate to the multinomial distribution $P(w_{-i}|z_{-i}, \beta_k)$.[**?**] Thus we can derive Equation 19 as

$$\begin{aligned}
P(\beta_k|z_{-i}, w_{-i}) &\propto P(w_{-i}|z_{-i}, \beta_k)P(\beta_k|z_{-i}) \\
&\propto \prod_{z_{(j\neq i)}=k} \prod_{v=1}^{V} \beta_{k,v} \cdot \prod_{v=1}^{V} \beta_{k,v}^{\eta-1} \\
&= \prod_{v=1}^{V} \beta_{k,v}^{n_{k,-i}^{(w_i)}} \cdot \prod_{v=1}^{V} \beta_{k,v}^{\eta-1} \\
&= \prod_{v=1}^{V} \beta_{k,v}^{n_{k,-i}^{(w_i)}+\eta-1}
\end{aligned}$$

where $n_{k,-i}^{(w_i)}$ is the number of times that the word w_i is assigned to topic k besides the current word. From this derivation we can find that the posterior distribution $P(\beta_k|z_{-i}, w_{-i})$ is $Dir_V(n_{k,-i}^{(w_i)} + \eta)$.

Looking back to Equation 18, it can be written as

$$\begin{aligned}
P(w_i|z_i = k, z_{-i}, w_{-i}) &= \int P(w_i|z_i = k, \beta_k)P(\beta_k|z_{-i}, w_{-i})d\beta_k \\
&= \int \beta_{k,w_i} P(\beta_k|z_{-i}, w_{-i})d\beta_k \\
&= E(\beta_k|z_{-i}, w_{-i}) \\
&= \frac{n_{k,-i}^{(w_i)} + \eta}{n_{k,-i}^{(\cdot)} + V\eta}
\end{aligned}$$

where $n_{k,-i}^{(\cdot)}$ is the total number of words assigned to topic k besides the current word.

Now we look at the term $P(z_i = k|z_{-i})$. Similar to the way for obtaining $P(w_i|z_i = k, z_{-i}, w_{-i})$, we can find that $P(\theta_d|z_{-i})$ is the Dirichlet distribution with parameter $n_{k,-i}^{(d)}+\alpha$ where $n_{k,-i}^{(d)}$ is the number of words assigned to topic k besides the current one in document d. Then,

$$\begin{aligned} P(z_i = k|z_{-i}) &= \int P(z_i = k|\theta_d)P(\theta_d|z_{-i})d\theta_d \\ &= \int \theta_{d,k}P(\theta_d|z_{-i})d\theta_d \\ &= E(\theta_d|z_{-i}) \\ &= \frac{n_{k,-i}^{(d)} + \alpha}{n_{\cdot,-i}^{(d)} + K\alpha}. \end{aligned}$$

where $n_{\cdot,-i}^{(d)}$ is the number of words in the document d besides the current one. Now combining the results above, we can rewrite Equation 17 as

$$P(z_i = k|z_{-i}, w) \propto \frac{n_{k,-i}^{(w_i)} + \eta}{n_{k,-i}^{(\cdot)} + V\eta} \cdot \frac{n_{k,-i}^{(d)} + \alpha}{n_{\cdot,-i}^{(d)} + K\alpha}. \tag{20}$$

The derivation of (20) follows the guidance of Griffiths work.[**?**]

By the definition of $\beta_{k,v}$ and $\theta_{d,t}$, we can estimate these two parameters by the posterior distribution such that

$$\hat{\beta}_{k,v} = \frac{n_{k,-i}^{(w_i)} + \eta}{n_{k,-i}^{(\cdot)} + V\eta} \tag{21}$$

$$\hat{\theta}_{d,k} = \frac{n_{k,-i}^{(d)} + \alpha}{n_{\cdot,-i}^{(d)} + K\alpha} \tag{22}$$

Knowing the method of evaluating the posterior distribution, we can figure out the Gibbs sampling scheme for the LDA model. Suppose we are estimating the topic mixture of document $d \in D$ such that it contains N_d words in total. The latent topics are indexed by $\{1, ..., K\}$, the vocabulary is indexed by $\{1, ..., V\}$:

1. Set $i = 1$, pick two positive integers α, η uniformly from the interval $[0, 1]$.
2. Pick word $w_i \in d$, $n_d = N_d - 1$.
3. Set $Z^{(0)} \sim Multinomial(N_d, \frac{1}{K}, ..., \frac{1}{K})$.
4. Sample $z_i^{(0)}$ from $Z^{(0)}$.
5. Update counts $n_{z_i^{(0)}} = n_{z_i^{(0)}} + 1$, $n_{z_i^{(0)},w_i} = n_{z_i^{(0)},w_i} + 1$, $n_{z_i^{(0)}} = n_{z_i^{(0)}} + 1$.
6. For k=1,...,K, $p(z_i = k|z_{-i}, w) = \frac{n_{d,k}+\alpha}{n_d+K\alpha} \frac{n_{k,w_i}+\eta}{n_k+V\eta}$, call this as the distribution of random variable $Z^{(1)}$ where $w = \{w_1, ..., w_{N_d}\}$.
7. Sample $z_i^{(1)}$ from $Z^{(1)}$.
8. Update counts $n_{d,z_i^{(0)}} = n_{d,z_i^{(0)}} - 1$, $n_{z_i^{(0)},w_i} = n_{z_i^{(0)},w_i} - 1$, $n_{z_i^{(0)}} = n_{z_i^{(0)}} - 1$,$n_{d,z_i^{(1)}} = n_{d,z_i^{(1)}} + 1$, $n_{z_i^{(1)},w_i} = n_{z_i^{(1)},w_i} + 1$, $n_{z_i^{(1)}} = n_{z_i^{(1)}} + 1$.
9. $Z^{(0)} = Z^{(1)}$.
10. If $i + 1 \leq N_d$, $i = i + 1$, go back to step 4; otherwise, go to the next document.

Now let's look at an example of applying Gibbs sampler to evaluate topic mixtures of a corpus.

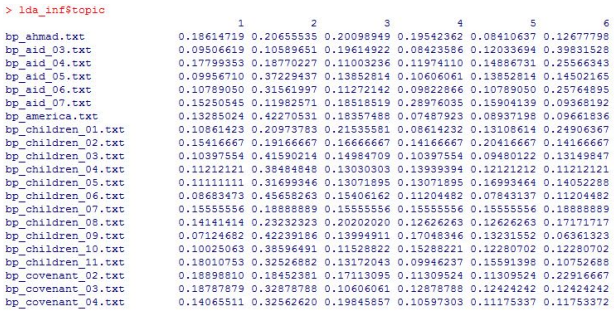
```
> terms(lda, 20) #extract the 20 most likely terms for each topic
      Topic 1 Topic 2   Topic 3      Topic 4 Topic 5   Topic 6
 [1,] "thy"   "the"     "the"        "thy"   "the"     "thee"
 [2,] "and"   "and"     "thou"       "the"   "thy"     "thou"
 [3,] "the"   "thy"     "lord"       "them"  "and"     "god"
 [4,] "hath"  "with"    "and"        "their" "that"    "and"
 [5,] "who"   "thi"     "most"       "have"  "thee"    "art"
 [6,] "name"  "mai"     "world"      "that"  "thine"   "all"
 [7,] "him"   "art"     "thee"       "for"   "glori"   "whose"
 [8,] "hast"  "verili"  "all"        "from"  "all"     "thing"
 [9,] "thee"  "make"    "allgloriou" "thei"  "can"     "from"
[10,] "that"  "lord"    "call"       "god"   "have"    "who"
[11,] "which" "upon"    "exalt"      "love"  "mai"     "power"
[12,] "have"  "love"    "god"        "and"   "through" "but"
[13,] "been"  "merci"   "which"      "what"  "which"   "there"
[14,] "with"  "our"     "glorifi"    "everi" "except"  "face"
[15,] "didst" "kingdom" "abid"       "heart" "unto"    "will"
[16,] "from"  "bounti"  "bounti"     "thou"  "creat"   "unto"
[17,] "most"  "grace"   "great"      "not"   "own"     "servant"
[18,] "caus"  "light"   "from"       "such"  "ani"     "for"
[19,] "turn"  "grant"   "upon"       "soul"  "with"    "grace"
[20,] "unto"  "forgiv"  "hold"       "peopl" "prais"   "merci"
```

Figure 6: Result from R of the 20 most likely words for each topic

Example 3. In this example we discover the topic mixtures of prayer texts. Our corpus contains 346 bible scriptures on prayer. According to R, the texts have 2964 distinct vocabulary items in total. In other words, $V = 2964$. Since the scale of our corpus is not very large, we choose the number of topic as 6. For the programming part, we apply the R package "topicmodels". The R implementation can be found in Appendix C.

After running the algorithm in R, we pick the top 20 most likely terms for each topic. The result is in Figure 6. We can see words such as "the", "and" always being chosen as top choices by the algorithm, however, these vocabulary are not meaningful to us. This is a shortcoming of using Gibbs sampler here. To receive a better image of what does each topic talk about, we can look at more likely terms for each topic.

Figure 7 shows us the posterior distribution of topic mixture for each document. The sum of probability of each row is 1. According to the result, the probabilities of topics for many documents are really close. However we have some documents that that has larger probability assigned to one or two topics (by large, we mean the probability is more than 0.4). If we look at these documents we may discover what are the topics with large probability about.



```
> lda_inf$topic
                          1          2          3          4          5          6
bp_ahmad.txt      0.18614719 0.20655535 0.20098949 0.19542362 0.08410637 0.12677798
bp_aid_03.txt     0.09506619 0.10589651 0.19614922 0.08423586 0.12033694 0.39831528
bp_aid_04.txt     0.17799353 0.18770227 0.11003236 0.11974110 0.14886731 0.25566343
bp_aid_05.txt     0.09956710 0.37229437 0.13852814 0.10606061 0.13852814 0.14502165
bp_aid_06.txt     0.10789050 0.31561997 0.11272142 0.09822866 0.10789050 0.25764895
bp_aid_07.txt     0.15250545 0.11982571 0.18518519 0.28976035 0.15904139 0.09368192
bp_america.txt    0.13285024 0.42270531 0.18357488 0.07487923 0.08937198 0.09661836
bp_children_01.txt 0.10861423 0.20973783 0.21535581 0.08614232 0.13108614 0.24906367
bp_children_02.txt 0.15416667 0.19166667 0.16666667 0.14166667 0.20416667 0.14166667
bp_children_03.txt 0.10397554 0.41590214 0.14984709 0.10397554 0.09480122 0.13149847
bp_children_04.txt 0.11212121 0.38484848 0.13030303 0.13939394 0.12121212 0.11212121
bp_children_05.txt 0.11111111 0.31699346 0.13071895 0.13071895 0.16993464 0.14052288
bp_children_06.txt 0.08683473 0.45658263 0.15406162 0.11204482 0.07843137 0.11204482
bp_children_07.txt 0.15555556 0.18888889 0.15555556 0.15555556 0.15555556 0.18888889
bp_children_08.txt 0.14141414 0.23232323 0.20202020 0.12626263 0.12626263 0.17171717
bp_children_09.txt 0.07124682 0.42239186 0.13994911 0.17048346 0.13231552 0.06361323
bp_children_10.txt 0.10025063 0.38596491 0.11528822 0.15288221 0.12280702 0.12280702
bp_children_11.txt 0.18010753 0.32526882 0.13172043 0.09946237 0.15591398 0.10752688
bp_covenant_02.txt 0.18898810 0.18452381 0.17113095 0.11309524 0.11309524 0.22916667
bp_covenant_03.txt 0.18787879 0.32878788 0.10606061 0.12878788 0.12424242 0.12424242
bp_covenant_04.txt 0.14065511 0.32562620 0.19845857 0.10597303 0.11175337 0.11753372
```

Figure 7: The probability of topics for each document (partial results are shown)

5 Conclusion

Through the paper we find that Gibbs sampling is a convenient tool for drawing samples and finding distributions when the target distribution is not available or very hard to compute. However we also discover that it may not always give us satisfying results in every circumstances (think about what we met in Example 3). Our explanation about Gibbs sampler is theoretical and all the provided examples aim either to give readers concrete images of the algorithm's implementation or to show certain properties of the sampler. In the real life, people are also focus on the efficiency of the method. For example, Schervish and Carlin [**?**], Cowles and Carlin[**?**] talk about the convergence diagnostics of the Markov Chain in the algorithm. Readers who are interested in this topic may do further research based on these previous works.

6 Acknowledgement

I would like to express my sincere gratitude to my thesis supervisor and major advisor, Professor Lynne Butler, for her continuous support to my thesis research and mathematics study. I appreciate her vast knowledge and

enlightening ideas in mathematics and statistics. Her suggestions always help me to overcome difficulties that I meet through the thesis writing. Our discussions about future study and life choice are also unforgettable to me. I would like to thank Professor David Lippel for his advice to my thesis. I must thank the Mathematics Department of Haverford College for providing enormous resources and help to my major study and thesis writing. Last but not least, I would also like to thank my lovely family for their continuous support through my entire life.

A

R code of Example 1

```
# Bivariate Normal with mean=0, covariance=rho
rbvn<function (n, rho)
{
        x < rnorm(n, 0, 1)
        y < rnorm(n, rho * x, sqrt(1    rho^2))
        cbind(x, y)
}
bvn<rbvn(1000,0.8)
par(mfrow=c(3,2))
plot(bvn,col=1:10000)
plot(bvn,xlab="X",ylab="Y",xlim=c(4,4),ylim=c(4,4),
   col="brown",pch=20)
plot(ts(bvn[,1]))
plot(ts(bvn[,2]))
hist(bvn[,1],40hhe)
hist(bvn[,2],40)
par(mfrow=c(1,1))

#Estimation by gibbs sampler
gibbs<function (n, rho, a, b)
{
        mat < matrix(ncol = 2, nrow = n)
        x < a
        y < b #the initial value is (a,b)
        mat[1, ] < c(x, y)
        for (i in 2:n) {
                x < rnorm(1, rho * y, sqrt(1    rho^2))
```

```
                y <- rnorm(1, rho * x, sqrt(1    rho^2))
                mat[i, ] <- c(x, y)
        }
        mat
}
bvn1=gibbs(1000,0.8,2.5,2.5)
bvn2=gibbs(1000,0.8, 2.5, 2.5)
bvn3=gibbs(1000,0.8, 2.5,2.5)
bvn4=gibbs(1000,0.8,2.5, 2.5)

colnames(bvn1)=c("X","Y")
df1=data.frame(bvn1)
plot(bvn1[1:10,],xlab="X",ylab="Y",xlim=c( 4,4),ylim=c
    ( 4,4),pch=19,col=1:10)
text(1,3.5,"Start Value")
arrows(1.2,3.3,2.5,2.5,length = 0.1, angle = 30, code =
     2)
lines(bvn1[1:10,])
par(new=TRUE)
plot(bvn2[1:10,],xlim=c( 4,4),ylim=c( 4,4),col=1:10,
    xaxt='n',yaxt='n', ann=FALSE)
lines(bvn2[1:10,])
par(new=TRUE)
plot(bvn3[1:10,],xlim=c( 4,4),ylim=c( 4,4),col=1:10,
    xaxt='n',yaxt='n', ann=FALSE)
lines(bvn3[1:10,])
par(new=TRUE)
plot(bvn4[1:10,],xlim=c( 4,4),ylim=c( 4,4),col=1:10,
    xaxt='n',yaxt='n', ann=FALSE)
lines(bvn4[1:10,])
```

```
points(x,y,pch=19)

x=c(2.5,2.5, 2.5, 2.5)
y=c(2.5, 2.5,2.5, 2.5)
plot(bvn1,type="l",xlab="X",ylab="Y",xlim=c(4,4),ylim=
    c(4,4),col='red',pch=20)
#lines(bvn1,col="red")
par(new=TRUE)
plot(bvn2,type="l",xlim=c(4,4),ylim=c(4,4),col='blue'
    ,xaxt='n',yaxt='n',pch=20, ann=FALSE)
#lines(bvn2,col="blue")
par(new=TRUE)
plot(bvn3,type="l",xlim=c(4,4),ylim=c(4,4),col='green
    ',xaxt='n',yaxt='n',pch=20, ann=FALSE)
#lines(bvn3,col="green")
par(new=TRUE)
plot(bvn4,type="l",xlim=c(4,4),ylim=c(4,4),col='
    purple',xaxt='n',yaxt='n',pch=20, ann=FALSE)
#lines(bvn4,col="purple")
points(x,y,pch=19)
text(2.5,2.8,"A")
text(2.8,2.8,"B")
text(3, 2.5,"C")
text(2.8, 2.5,"D")

plot(bvn1[501:1000,],xlab="X",ylab="Y",xlim=c(4,4),
    ylim=c(4,4),col='red',pch=20)
#lines(bvn1)
par(new=TRUE)
```

```
plot(bvn2[501:1000,],xlim=c(4,4),ylim=c(4,4),col='
    blue',xaxt='n',yaxt='n',pch=20,   ann=FALSE)
#lines(bvn2)
par(new=TRUE)
plot(bvn3[501:1000,],xlim=c(4,4),ylim=c(4,4),col='
    green',xaxt='n',yaxt='n',pch=20,  ann=FALSE)
#lines(bvn3)
par(new=TRUE)
plot(bvn4[501:1000,],xlim=c(4,4),ylim=c(4,4),col='
    purple',xaxt='n',yaxt='n',pch=20,  ann=FALSE)
#lines(bvn4)

hist(bvn[,1],40)
hist(bvn[,2],40)
```

B

R code of Example 2

```
bodytemp=read.table("bodytemp.txt",header=T)
y=bodytemp$temp
ybar=mean(y); n=length(y)
#Get the 1000 samples from Gibbs sampler (we discard
    the first 4000 data for completing the burn in
    method)
Iterations=5000
mu0=98.6; s0=0.5; a0=0.001; b0=0.001
theta = matrix(nrow=Iterations, ncol=2)
mu=0; tau=2; temp.s=sqrt(1/tau)
for (t in 1:Iterations){
     w= s0^2/( temp.s^2/n+ s0^2 )
     m = w*ybar + (1 w)*mu0
     s = sqrt( w/n ) * temp.s
     mu = rnorm( 1, m, s )
     a = a0 + 0.5*n
     b = b0 + 0.5 * sum( (y mu)^2 )
     tau = rgamma( 1, a, b )
     temp.s = sqrt(1/tau)
     theta[t,]=c( mu, temp.s)
}

mcmc.output=theta
apply(mcmc.output[ (1:1000) ,],2,mean)
#compare to true value: 98.25, 0.542
apply(mcmc.output[ (1:1000) ,],2,sd)
#compare to true value: 0.06456, 0.06826
```

```
#Making the histogram of mean and variance
histdata=function(mu0,s0,a0,b0,it1,it2){
mudata=rep(0,it2)
sdata=rep(0,it2)
for (i in 1:it2){
          for (t in 1:it1){
         mu=mu0; tau=2; temp.s=sqrt(1/tau)
         w= s0^2/( temp.s^2/n+ s0^2 )
         m = w*ybar + (1 w)*mu0
         s = sqrt( w/n ) * temp.s
         mu = rnorm( 1, m, s )
         a = a0 + 0.5*n
         b = b0 + 0.5 * sum( (y mu)^2 )
         tau = rgamma( 1, a, b )
         temp.s = sqrt(1/tau)
        }
   mudata[i]= mu
   sdata[i]= temp.s
}
return(list(mudata=mudata,sdata=sdata))
}

sample=histdata(98.6,0.5,0.001,0.001,5000,2000)
hist(sample$mudata,30,freq=T,xlab="Data sample of mu",
    main="Histogram of Mu")
mean(sample$mudata)
var(sample$mudata)
```

C

R code of Example 3

```
library(tm)
library(topicmodels)
files = DirSource(directory = "E:/Documents/Senior Thesis/R work/PrayerCorpusPartOne/",encoding ="latin1" )
corpus = VCorpus(x=files)
summary(corpus)
corpus = tm_map(corpus,removePunctuation)
corpus = tm_map(corpus,stripWhitespace)
corpus = tm_map(corpus,removePunctuation)
matrix_terms = DocumentTermMatrix(corpus) #Transform the data to a document term matrix
lda=LDA(matrix_terms, 6, method = "Gibbs") #We obtain the lda gibbs model for the corpus
lda_inf = posterior(lda, matrix_terms)
#Determine the posterior probabilities of the topics for each document and of the terms for each topic for a fitted  topic model.
lda_inf$topic $the distribution over all topics for each document
terms(lda, 20) #extract the 20 most likely terms for each topic
topics(lda, 6) #Sort the most likely topics for each document.
```

References

[1] A.Gelman,*Bayesian Data Analysis Third Edition*. CRC Press, Taylor & Francis Group, Boca Raton, USA, 2014.

[2] S.Brook and A.Gelman,*Handbook of Markov Chain Monte Carlo*. CRC Press, Taylor & Francis Group, Boca Raton, USA, 2011.

[3] C.P.Robert and G.Casella,*Introducing Monte Carlo Methods in R*. Springer, New York, 2010.

[4] G.Casella and E. I. George,*Explaining the Gibbs Sampler*. The American Statistician, Vol. 46, No. 3 (Aug., 1992), 167-174.

[5] J. Besag, *Spatial Interaction and the Statistical Analysis of Lattice Systems*. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 36, No. 2 (1974),192-236.

[6] M. J. Schervish and B. P. Carlin,*On the Convergence of Successive Substitution Sampling*. Journal of Computational and Graphical Statistics, Vol. 1, No. 2 (Jun., 1992),111-127.

[7] M.K. Cowles and B. P. Carlin,*Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review*. Journal of the American Statistical Association, Vol. 91, No. 434 (Jun., 1996),883-904.

[8] L.Tierney,*Markov Chains for Exploring Posterior Distributions*. Annals of Statistics,Vol.22 No.4 (Dec., 1994),1701-1728.

[9] U. E. Makov, A. F. M. Smith and Y.-H. Liu,*Bayesian Methods in Actuarial Science*. Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 45, No. 4 (1996),503-515.

[10] D. Blei *Probabilistic Topic Models* . Communications of the ACM, Vol. 55 No. 4(Apr., 2012), 77-84.

[11] D. Blei, A. Ng, M. Jordan and J. Lafferty *Latent Dirichlet Allocation*. Journal of Machine Learning, 3:993-1022, 2003.

[12] P.A. Mackowiak, S.S. Wasserman, M.M. Levine *A Critical Appraisal of 98.6F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich*. The Journal of the American Medical Association, 268 (12): 15781580, 1992.

[13] T.Minka and J. Lafferty,*Expectation-Propagation For the Generative Aspect Model*. In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, 2002

[14] T. Griffiths and M. Steyvers,*A Probabilistic Approach to Semantic Representation*. In Proceedings of the 24th annual conference of the cognitive science society, pp. 381-386. 2002.

[15] T. Griffiths and M. Steyvers,*Finding Scientific Topics*. Proceedings of the National Academy of Sciences 101, no. suppl 1 (2004): 5228-5235.

[16] T.Griffiths, *Gibbs Sampling In the Generative model of Latent Dirichlet Allocation*. Stanford University, 2002.

[17] D.A. Levin, *Markov Chains and Mixing Times*. American Mathematical Society, Providence, USA, 2009

[18] F.P. Kelly, *Reversibility and Stochastic Networks*. Cambridge University Press. Cambridge, UK, 2011