

Bayesian Conjugate Analysis for Transition Probabilities of Non-Homogeneous Markov Chain: A Survey

Minje Sung^{1,a}

^aSchool of Business, Ajou University, Korea

Abstract

The present study surveys Bayesian modeling structure for inferences about transition probabilities of Markov chain. The motivation of the study came from the data that shows transitional behaviors of emotionally disturbed children undergoing residential treatment program. Dirichlet distribution was used as prior for the multinomial distribution. The analysis with real data was implemented in WinBUGS programming environment. The performance of the model was compared to that of alternative approaches.

Keywords: Bayesian approach, transition probability, Markov chain.

1. Introduction

We will use the Markov chain in discrete time to analyze the transient states of emotionally disturbed children and adolescents undergoing residential treatment. By understanding the nature of the transitions of the states, we will be able to predict and exercise some control over the future states of the patients. Since the patients' states changes over time with uncertainty, the future states are not deterministic, and we can handle the uncertainty of the states by assessing probabilities to the process. The statistical inferences about the patients' transitional behavior are made under the Bayesian framework where the probability statements about the parameter are conditional on the observed data. The patients' transitional states are described by a Markov Chain where the current patients' states only depend on the immediate past not the whole history. At each time point, the patients' states have multinomial distribution. Using Dirichlet as a conjugate prior, the posterior distributions are computed. The posterior distributions are represented by simulations using Markov Chain Monte Carlo (MCMC) method.

The early efforts to make inferences on the transition probabilities of a Markov chain was made in Anderson and Goodman (1957), and Billingsley (1961) where the inferences are made using maximum likelihood methods. Duncan and Lin (1972) considered models for a Markov chain having stochastic entry and exits. In other words, the authors relaxed the assumption that the number of observed individuals over time is constant and allowed for changes as a result of units entering or leaving the system. Such models, especially those allowing stochastic exits, are of interest in the analysis of data from psychiatric treatment programs where patients may leave the treatment program for different reasons. However they were non Bayesian approaches. A Bayesian analysis of the Markov chains using Dirichlet conjugate prior distribution is presented in Lee *et al.* (1970) but no explicit model is considered to describe time dependent transition probabilities over time. Meshkani and Billard (1992) developed empirical Bayes estimates of transition probabilities for homogeneous chains and extended

¹ Associate Professor, School of Business, Ajou University, San 5 Woncheon-Dong Yeongtong-Gu, Suwon 443-749, Korea. E-mail: sungmj@ajou.ac.kr

these to nonhomogeneous Markov chains by viewing the problem as a parametric Bayes problem in the sense of Morris (1983). However, the empirical Bayes approach is not a Bayesian approach. The literature does not show much of Bayesian studies for Markov chains. Sung *et al.* (2007) developed Bayesian models for non-homogeneous Markov chains by using logistic type of models with covariates. However, Markov chain models sometimes do not have covariate information of subjects, and just show aggregate transition patterns. In the present study, we propose a Bayesian approach for Markov chains without covariate information.

In Bayesian paradigm, inferences about unknown parameters in statistical models are made using posterior samples mostly drawn from Monte Carlo methods. For the choice of certain class of prior distributions, the posterior distribution follows the same parametric form as the prior distribution. This property is referred to as conjugacy and the prior distribution is then called a conjugate prior distribution for the particular likelihood. A formal treatment of conjugate Bayesian analysis is given in Bernardo and Smith (1994, Section 5.2).

While not many but some rigorous and novel Bayesian approaches for nonhomogeneous Markov chains were proposed in literature (for example, Sung *et al.*, 2007), the Bayesian analysis in the present study adds a practical contribution to the literature in that the analysis of present study is straightforward but comprehensive conjugate Bayesian way, and practically the present model is much more accessible without having the barrier of having to understand the complexity of models for implementation. The performance of present model was compared to the alternative approach in the literature. Our primary concern in this study is to review a straightforward Bayesian conjugate analysis for Markov chain, and examine from a practical point of view how fully Bayesian method can give insights about the patients' states to practitioners. A comprehensive description of Bayesian conjugate analysis is discussed in Section 2. Section 3 has a demonstration of model implementation using real data, and the model performance is compared to alternative approaches in the literature. Section 4 summarizes the findings of present study.

2. The Markov Probability Model

The type of stochastic process with which we will be concerned has a finite number of possible outcomes S_j ($j = 1, \dots, J$) where a discrete random variable X_t ($t = 0, 1, \dots, T$) takes at a finite number of equidistant time points. The probability distribution, $P(X_0 = S_j(0), X_1 = S_j(1), \dots, X_T = S_j(T))$, for an individual would be $\prod_{t=0}^T P(X_t = S_j(t))$ if the states were independent in each stage. If we assume first order dependence where the probability distribution of a state of a given stage only on the state of the immediately preceding stage, then the process has the Markov property, and the Markov chain can be written as

$$P(X_0 = S_i, X_1 = S_j, \dots, X_T = S_l) = P(X_0 = S_i)P(X_1 = S_j | X_0 = S_i) \cdots P(X_T = S_l | X_{T-1} = S_k). \quad (2.1)$$

For the nonhomogeneous Markov Chains, it is assumed that if $X_{t-1} = S_i$ and $X_t = S_j$, then $P(X_t = S_j | X_{t-1} = S_i) = P_{ij}(t)$. And, N_{ij} denotes the number of individuals such that $X_{t-1} = S_i$ and $X_t = S_j$. The transition probability matrix ($I \times J$) has the following properties.

$$0 \leq p_{ij}(t) \leq 1 \text{ and } \sum_j p_{ij}(t) = 1, \quad \text{for } i = 1, \dots, I, j = 1, \dots, J, t = 1, \dots, T.$$

The number of subjects who made transitions from S_i at time $t - 1$ to S_j at time t is denoted by the random variables, $N_{ij}(t)$, and the random vector $\tilde{N}_i(t) = (N_{i1}(t), \dots, N_{iJ}(t))$ has a multinomial

distribution with parameters $n_i(t)$ and $p_{ij}(t)$'s. Then the pdf of the multinomial distribution, can be defined by

$$f_i(\tilde{N}_i(t) | n_i(t), p_{ij}(t)) = P(N_{i1}(t) = n_{i1}(t), \dots, N_{iJ}(t) = n_{iJ}(t))$$

and can be written as

$$f_{i(t)}(n_{i1}(t), \dots, n_{iJ}(t) | n_i(t), p_{ij}(t)) = \frac{n_i(t)!}{n_{i1}(t)! \dots n_{iJ}(t)!} p_{i1}(t)^{n_{i1}(t)} \dots p_{iJ}(t)^{n_{iJ}(t)}, \quad (2.2)$$

where $n_i(t) = \sum_{j=1}^J n_{ij}(t)$. We are assuming that transitions from state are independent of transitions from state k . For the rows of the transition matrix, the multinomial distribution for a give row i is independent of the other rows and this means that $\tilde{N}_i(t)$ is independent of \tilde{N}_k where $i \neq k$.

When the joint pdf $f_i(\tilde{N}_i(t) | n_i(t), p_{ij}(t))$ is regarded as a function of the parameters for a given vector $\tilde{N}_i(t)$, it is called the likelihood function. If we are to estimate the parameters, we would wish to find the parameters for which the probability of obtaining the actual observed vector $\tilde{n}_i(t) = (n_{i1}(t), \dots, n_{iJ}(t))$ is very high. And we will use this value as the estimates of the parameters. For this reasoning, we will use the estimates of the parameters for which the likelihood function is maximum. Let $n_i(t) = \sum_j n_{ij}(t)$, and assume that each row is independent of the other rows. Then the likelihood function is

$$\begin{aligned} l(\tilde{p}_i(t) | \tilde{n}_i(t)) &= \prod_{i=1}^I \left[\frac{n_i(t)!}{\prod_{j=1}^J n_{ij}(t)!} \prod_{j=1}^J p_{ij}(t)^{n_{ij}(t)} \right] \\ &\propto \prod_{i=1}^I \prod_{j=1}^J p_{ij}(t)^{n_{ij}(t)}. \end{aligned} \quad (2.3)$$

In the Bayesian approach, it is assumed an investigator's information or uncertainty about a parameter vector $\tilde{\theta}_i = (p_{i1}, \dots, p_{iJ})$ for a given row i can be summarized in a prior probability density function $g_i(\tilde{\theta}_i)$. By the use of Bayes' theorem, this information can be combined with the sample density function $f_i(\tilde{N}_i | \tilde{\theta}_i)$, where \tilde{N}_i denotes a vector of sample observations in row i . And the posterior pdf, $g_i(\tilde{\theta}_i | \tilde{N}_i)$, can be used to make inferences about the parameters. The Bayes' theorem is as follows. Let $f_i(\tilde{N}_i | \tilde{\theta}_i)$ denote the joint probability density function of \tilde{N}_i and $\tilde{\theta}_i$. Then we have

$$f_i(\tilde{N}_i, \tilde{\theta}_i) = l_i(\tilde{N}_i | \tilde{\theta}_i) g_i(\tilde{\theta}_i) = g_i(\tilde{\theta}_i | \tilde{N}_i) g_i(\tilde{N}_i). \quad (2.4)$$

Thus, the posterior pdf for the parameter vector $\tilde{\theta}_i$, given the sample information \tilde{N}_i , is

$$g_i(\tilde{\theta}_i | \tilde{N}_i) \propto g_i(\tilde{\theta}_i) l_i(\tilde{N}_i | \tilde{\theta}_i), \quad (2.5)$$

where $l_i(\tilde{N}_i | \tilde{\theta}_i)$ is the likelihood function.

Assuming the Dirichlet as a prior distribution and independence of rows, the posterior pdf is in the form of a product of Dirichlet pdf's with parameters $n_{ij}(t) + a_{ij}(t-1)$ as

$$g_i(\tilde{p}_i(t) | \tilde{n}_i(t)) \propto \prod_{i=1}^I \left\{ \frac{\Gamma(\sum_{j=1}^J (n_{ij}(t) + a_{ij}(t-1)))}{\prod_{j=1}^J \Gamma(n_{ij}(t) + a_{ij}(t-1))} \prod_{j=1}^J p_{ij}(t)^{n_{ij}(t) + a_{ij}(t-1) - 1} \right\}, \quad (2.6)$$

where Γ denotes the gamma function and the a_{ij} 's are positive parameters which can be thought as prior sample sizes.

Here, posterior independence is preserved. Thus, each row of the matrix is a Dirichlet distribution.

Further, the marginal posterior pdf for a single transition probability is in the standard beta form with mean, variance, and covariance given by

$$E(p_{ij}(t) | n_i(t)) = \frac{n_{ij}(t) + a_{ij}(t-1)}{c_i(t)},$$

$$\text{Var}(p_{ij}(t) | n_i(t)) = \frac{(n_{ij}(t) + a_{ij}(t-1))(c_i(t) - n_{ij}(t) - a_{ij}(t-1))}{c_i(t)^2(c_i(t) + 1)},$$

where $c_i(t) = \sum_{j=1}^J (n_{ij}(t) + a_{ij}(t-1))$, and $n_i(t) = \sum_{j=1}^J n_{ij}(t)$.

It can be shown that the maximum likelihood estimator (MLE) of $p_{ij}(t)$ to be $n_{ij}(t) / \sum_{j=1}^J n_{ij}(t)$, then the posterior mean can be written as the weighted average of the MLE and the prior mean,

$$w \frac{n_{ij}(t)}{\sum_{j=1}^J n_{ij}(t)} + (1-w) \frac{a_{ij}(t-1)}{\sum_{j=1}^J a_{ij}(t-1)},$$

where $1 \leq w \leq 1$. If the size of the sample becomes large, then more weight will be given to the sample mean. Thus the posterior mean will be getting closer to the MLE as the sample size gets larger.

Sometimes it is more realistic and practical to make inferences about the now unknown but eventually observable $\tilde{n}_i(t+1)$. Thus, the posterior predictive distribution given row i can be written as

$$P_i(\tilde{n}_i(t+1) | \tilde{n}_i(t)) = \int f_i(\tilde{n}_i(t+1) | \tilde{p}_i(t), \tilde{n}_i(t)) g_i(\tilde{p}_i(t) | \tilde{n}_i(t)) d\tilde{p}_i(t). \quad (2.7)$$

Then, using Dirichlet distribution as a natural conjugate prior for the multinomial distribution with J categories, the posterior predictive density of $\tilde{n}_i(t+1) = (n_{i1}(t+1), \dots, n_{iJ}(t+1))$ is

$$P_i(\tilde{n}_i(t+1) | \tilde{n}_i(t))$$

$$= \int_0^1 \left\{ \frac{n_i(t+1)!}{\prod_{j=1}^J n_{ij}(t+1)!} \prod_{j=1}^J p_{ij}(t)^{n_{ij}(t+1)} \right\} \times \left\{ \frac{\Gamma(n_i(t) + a_i(t-1))}{\prod_{j=1}^J \Gamma(n_{ij}(t) + a_{ij}(t-1))} \prod_{j=1}^J p_{ij}(t)^{n_{ij}(t) + a_{ij}(t-1) - 1} \right\} dp(t)$$

$$= \frac{n_i(t+1)!}{\prod_{j=1}^J n_{ij}(t+1)!} \frac{\Gamma(n_{ij}(t) + a_{ij}(t-1))}{\prod_{j=1}^J \Gamma(n_{ij}(t) + a_{ij}(t-1))} \frac{\prod_{j=1}^J \Gamma(n_{ij}(t+1) + n_{ij}(t) + a_{ij}(t-1))}{\Gamma(n_i(t) + a_i(t-1))}, \quad (2.8)$$

where $n_i(t+1) = \sum_{j=1}^J n_{ij}(t+1)$, $n_i(t) = \sum_{j=1}^J n_{ij}(t)$, and $a_i(t-1) = \sum_{j=1}^J a_{ij}(t-1)$. The posterior predictive distribution in (2.8) is for a given row i and time $t+1$. We assume the subjects' states at a given time point is independent of each other, and the each row of the transition matrix is independent of the other rows. Thus, for I rows, the predictive distribution will be product over i .

3. Examples

3.1. Application to real data

The data are from the research conducted by Dr. N. Nhan (1998), where he collected the data at a psychiatric facility for children and adolescents, and used the data to describe how the states of the

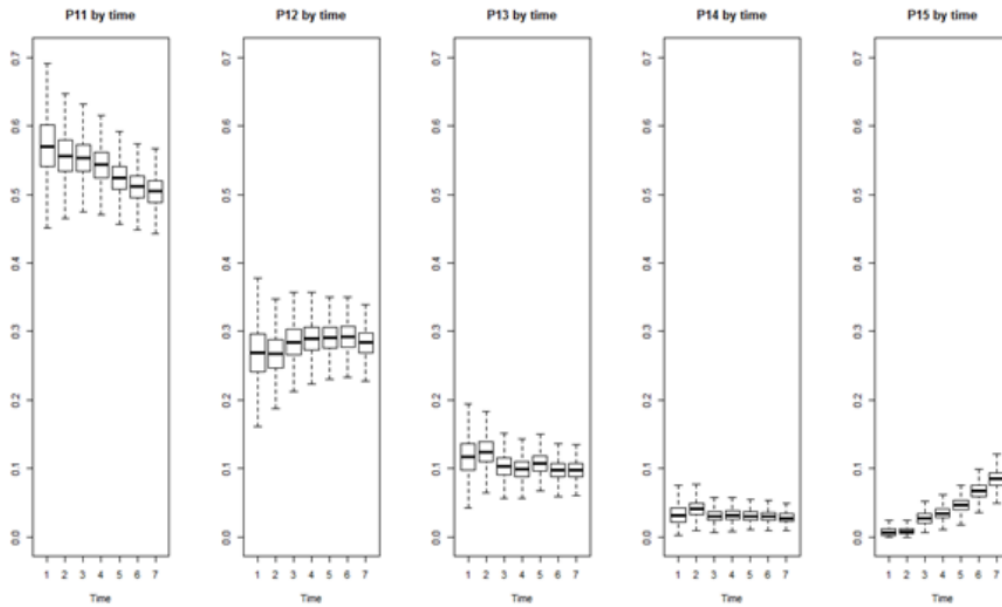


Figure 1: $p_{1j}(t)$ for $j = 1, \dots, 5, t = 1, \dots, 7$.

patients change during treatment as time passes in a non-Bayesian way. The treatment program at this facility is based on psychodynamic principles and is interdisciplinary in approach. The goal is to help them overcome whatever deficiencies in their personality and behavior so they could return to their family and community. We are interested in whether the planned treatment is effective.

The data were collected on 384 patients by treatment team members during the period of their inpatient treatment. Since patients are leaving at different time points, the number of patients is decreasing over time. There are four states with 1 being the lowest and 4 being the highest. In addition, there is 5th state, which is an absorbing state for the patients who are leaving the facility. The period starts from 30 days after their admission (time = 0) and approximately every three months thereafter. We will use the data up to the 7th period.

We assumed flat priors by setting $a_{ij}(0) = 1$ for all i 's and j 's. We obtained plots of posterior distributions by simulation of 5,000 samples after 100,000 burn in with the WinBUGS program (Spiegelhalter *et al.*, 1996). The WinBUGS code is in Appendix. In Bayesian paradigm it is more desirable to make inferences using probability statements rather than using point estimates.

We illustrate the inferences about transition probabilities in Figure 1, which shows how transition probabilities from state 1 to state $j = 1, \dots, 5$ change over $t = 1, \dots, 7$. Here state $j = 5$ represents exit state. It is noted that $p_{11}(t)$ decreases with time indicating that the subjects are less and less staying in the same state once they are in state 1. $p_{12}(t)$ and $p_{15}(t)$ are gradually increasing with time whereas $p_{13}(t)$ and $p_{14}(t)$ do not change much. These are transitions moving to a better state making improvements. However the subjects are not likely to jump to 2 or 3 steps higher state, but more likely to make one step improvement. Also, we note that the probability of exiting gets higher with time which implies that the subjects are likely to leave the facility as time goes when they are in the lowest state.

It is possible to assess exit behavior of subjects by including the exit state ($j = 5$) in the transition

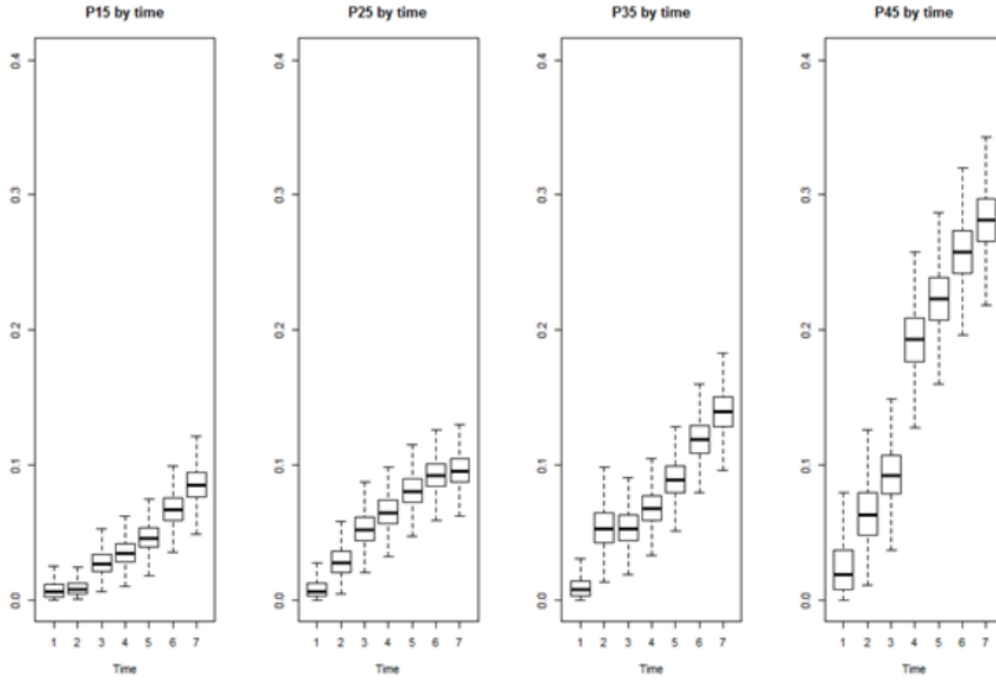


Figure 2: Exit behavior over time

matrix. Figure 2 shows posterior probabilities for exit behaviors making transition to state 5 (exit state) from states 1, 2, 3, and 4. Overall, the probability of exit increases with time from all states, and the subjects are more likely to leave the facility from higher states than from lower states. This kind of exit behavior implies that the subjects are leaving because their states have been improved moving to a higher state.

One of advantages of Bayesian methodology is that it allows us to make multiple comparisons of parameters. Inferences concerning two or more unknown quantities can also be made using the posterior distribution. For example, a comparison of any two transition probabilities $p_{ij}(t)$ and $p_{ik}(t)$ of the transition matrix can be made based on the posterior distribution

$$P(p_{ij}(t) - p_{ik}(t) > 0) = \int_{p_{ij}(t) - p_{ik}(t) > 0} P(p_{ij}(t) - p_{ik}(t) > 0) dp_{ij}(t) p_{ik}(t),$$

and the above can be approximated by Monte Carlo integral as

$$\frac{1}{G} \sum_{g=1}^G 1(p_{ij}^g(t) - p_{ik}^g(t) > 0),$$

where $\{p_{ij}^g(t); g = 1, \dots, G\}$ and $1(A)$ takes the value 1 if event A occurs and 0 otherwise. Figure 3 shows such multiple comparisons, which would be difficult to make in classical statistics. When $p_{45}(t)$ and $p_{35}(t)$ were compared, the distributions show positive differences especially for later times indicating higher probability of exit from state 4 than from state 3. The middle and right panels shows box plots to address the question of “Is it more likely to make one step improvement from a higher

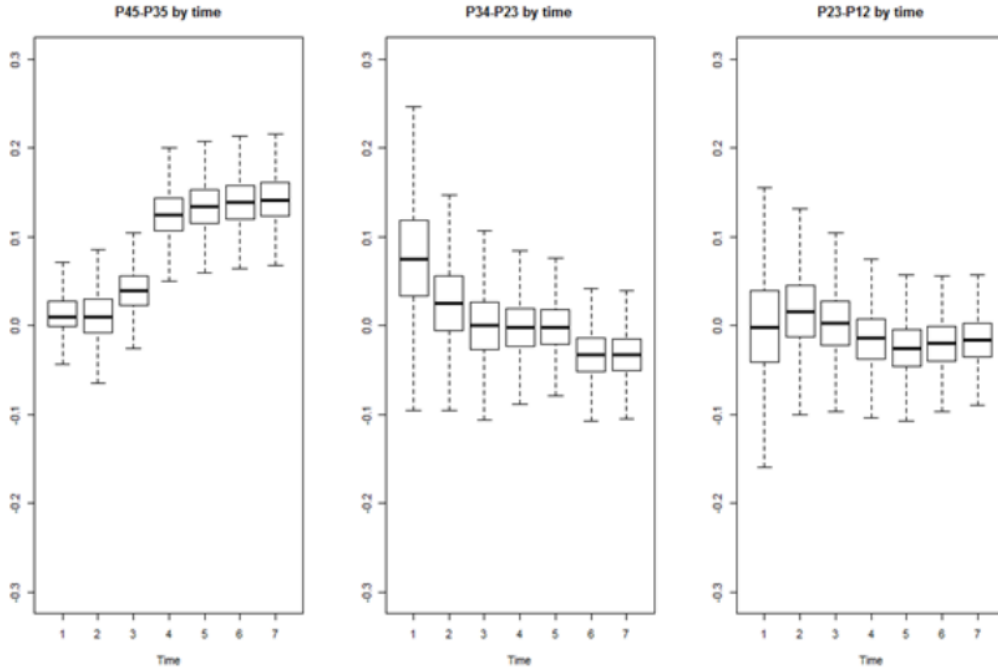


Figure 3: Comparisons of transition probabilities

state than from a lower state?” The question is partially supported as the box plots in the middle panel show mostly positive differences for early times, but this is not true in the right panel. This implies that subjects are more likely to make improvements from state 3 than from state 2 in early stage, but this chance decrease with time. When the subjects are in states 1 and 2, they are almost equally likely to make improvements.

Using the posterior predictive distribution in (2.8), the expected predictive transition matrix for $t = 8$ given all observed data up to $t = 7$ can be approximated from the Gibbs sampling as follows

$$E(\mathbf{N}(8) | \mathbf{N}(7)) = \begin{bmatrix} 50.5(45.9, 55.0) & 28.3(24.5, 32.4) & 9.8(7.2, 12.6) & 2.8(1.5, 4.6) & 8.5(6.2, 11.3) \\ 17.5(14.5, 20.8) & 37.0(33.1, 41.2) & 26.7(23.1, 30.5) & 9.2(7.0, 11.8) & 9.6(7.3, 12.1) \\ 4.8(3.1, 6.9) & 22.6(19.1, 26.6) & 35.2(30.9, 39.3) & 23.4(19.7, 27.2) & 13.9(11.0, 17.1) \\ 2.1(0.9, 3.8) & 7.7(5.2, 10.5) & 15.0(11.6, 18.8) & 47.0(42.1, 51.9) & 28.2(23.8, 32.9) \end{bmatrix},$$

where $\mathbf{N}(t) = (\tilde{n}_1(t), \dots, \tilde{n}_I(t))'$, and 95% credible intervals are in (). The Bayesian credible interval is a probability statement about $p_{ij}(t)$ and defined as

$$P(p_\star < p_{ij}(t) < p^\star | \tilde{n}_1(t), \dots, \tilde{n}_I(t)) = \int_{p_\star}^{p^\star} P(p_{ij}(t) | \tilde{n}_1(t), \dots, \tilde{n}_I(t)) dp_{ij}(t),$$

where $p_\star < p^\star$ are fixed values. If $P(p_\star < p_{ij}(t) < p^\star | \tilde{n}_1(t), \dots, \tilde{n}_I(t)) = (1 - \alpha)$, then (p_\star, p^\star) is referred to as $100(1 - \alpha)\%$ credible interval for $p_{ij}(t)$. In the above posterior predictive matrix for $\mathbf{N}(8)$, the 95% credible intervals were obtained from the Gibbs sampling, and can be interpreted analogous to classical confidence interval.

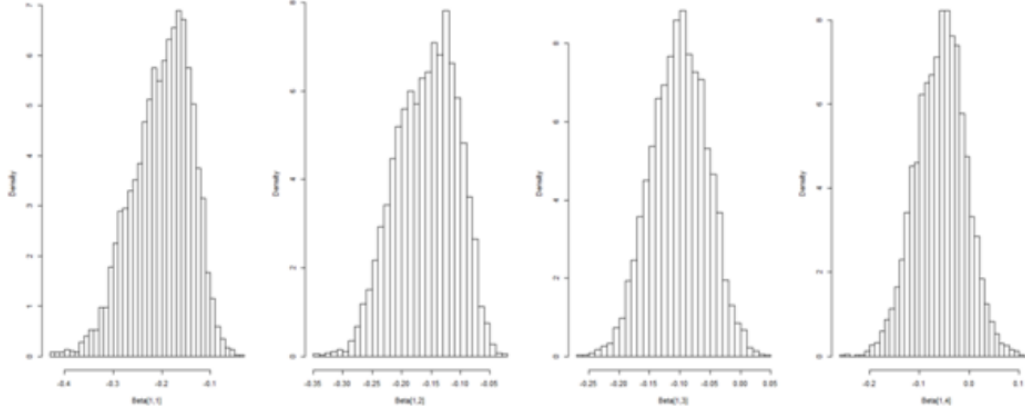


Figure 4: Posterior distributions of covariate effects from row $i = 1$

3.2. Comparison with alternative approaches

An alternative approach to describe the transitional behavior of non-homogeneous Markov chains was proposed by Sung *et al.* (2007) when covariate information was available. The present study used the same data but the aggregate transitional Markov chain without covariates was used whereas Sung *et al.* (2007) used individual Markov chain with covariates. Instead of relating the transition probabilities to the Dirichlet prior as in (2.6), Sung *et al.* (2007) used multinomial logistic regression model as follows

$$\text{logit}(p_{mijt}) = \log\left(\frac{p_{mijt}}{p_{miJt}}\right) = \gamma_{jt} + \gamma_{ijt} + \beta_{ij}z_{mt}, \quad (3.1)$$

where γ_{jt} is column and time specific component, γ_{ijt} is row, column, and time specific component, z_{mt} is a time dependent covariate, β_{ij} captures row and column specific covariate effects, and index m represents individual subjects for $m = 1, \dots, M$, $i = 1, \dots, I$, $j = 1, \dots, J - 1$. Column J was used as a baseline column for logit in (3.1). While more detailed prior and posterior analyses were described in Sung *et al.* (2007), the posterior distributions of covariate effects β_{ij} 's for $i = 1$ are represented in Figure 4 for illustration purpose. WinBUGS code for the model is in Appendix. For illustration purpose the age was used as a covariate. Posterior distributions of β_{1j} 's in Figure 4 shows mostly non-zero values except for β_{13} and β_{14} . Thus there is a significant age effect on the transitional probabilities. While not shown in the figure, other β_{ij} 's for $i = 2, 3, 4$, were considerably differently from zero. The posterior 95% credible intervals were $(-0.35, -0.13)$ for β_{11} , $(-0.28, -0.07)$ for β_{12} , $(-0.17, 0.04)$ for β_{13} , and $(-0.12, 0.09)$ for β_{14} .

With alternative modeling approaches having been considered, much of a concern is how good fit the model is. For complex models where the computation of Bayes Factors are challenging, Deviance Information Criterion(DIC) provides an efficient measure of model fit, which can easily be computed from posterior distribution of log likelihood (Spiegelhalter *et al.*, 2002). DIC is defined as

$$\text{DIC} = \bar{D} + p_D,$$

where $\bar{D} = E_{\theta|y}[-2 \ln f(y|\theta)]$ and $p_D = \bar{D} - D(\bar{\theta}) = E_{\theta|y}[D(\theta)] - D(E_{\theta|y}[\theta])$. DIC can be obtained by WinBUGS, or easily computed from the posterior samples obtained from Markov chain Monte

Table 1: Comparison of model fit using DIC

Model	\bar{D}	$D[\hat{\theta}]$	p_D	DIC
Homogeneous chain, present model	709.95	694.28	15.68	725.63
Nonhomogeneous chain, present model	648.95	613.80	35.16	684.11
Homogeneous chain, logistic regression	5070.47	5047.40	23.07	5093.54
Nonhomogeneous chain, logistic regression	4875.13	4830.87	44.26	4919.40

Carlo simulation. The better the model fit is, the larger the likelihood. Thus \bar{D} , which is -2 times log likelihood would get smaller for better models. p_D represents the effective number of parameters. Thus, a small value of DIC is preferred.

The performance of Bayesian conjugate multinomial model of (2.6) for nonhomogeneous Markov chains was compared to alternative Bayesian approach of Sung *et al.* (2007), which used the same data. Basically present study is a scale-down analysis of nonhomogeneous Markov chains compared to the study by Sung *et al.* (2007). However they are two different classes of models in that the present approach used straightforward Bayesian conjugate analysis without covariates whereas Sung *et al.* (2007) used regression type of modeling strategy to incorporate covariates to account for individual differences of transitional behaviors. When the performances of model fitting were compared, we also considered models for homogeneous Markov chains where $p_{ij}(t) = p_{ij}$ for all $t = 1, \dots, T$.

As summarized in Table 1, DIC favored models for nonhomogeneous chains over homogeneous chains for both present multinomial model and logistic regression type model of Sung *et al.* (2007). The model in the present study showed considerably smaller DIC value than the logistic regression type model. Even though the present model used aggregate data without covariate, and does not account for the heterogeneity of individual transitional pattern, the model took a huge advantage of having simpler structure than the logistic regression model.

4. Conclusion

We made inferences about the parameters through simulation technique using WinBUGS software. Density plots and statistics of the simulation results illustrate transitional behavior of the patients. Through the analyses on the posterior distributions, the researchers can shed some lights on the behavioral patterns of subjects while they are in the facility. The present study approach was implemented with real data to demonstrate some benefits of using Bayesian approach. By posterior simulation, it is possible to make inference about the difference of two transitional probabilities. Thus, it is possible to compare two probabilities directly from the simulation results.

The performance of model fit was compared to an alternative approach in the literature using the same data, and the DIC favored the model in the present study for simplicity. Future study may develop integrated Bayesian modeling framework of Dirichlet prior and regression type of models.

Appendix:

WinBUGS code for model of Section 3.1

```

model
{
for (t in 1:T){ #loop for time index
for (i in 1:I){ #loop for i-th row of Markov matrix for given t
x[t,i,1:J]~dmulti(p[t,i,1:J], M[t,i])
#for each row i, multinomial for J states where J=5 (including exit)

```

```
M[t,i]<-sum(x[t,i,1:J]) #total subjects for row i
p[t,i,1:J]~ddirch(alpha[t,i,1:J]) #Dirichlet prior
}}}}
```

WinBUGS code for an alternative model of Section 3.2

```
model
{
for (t in 2:T) {
for (m in 1:M){
y[t,x[t-1,m],m]~dcat(p[t,m,x[t-1,m],1:J])
#Individual transition data where m represents individual index
y[t,x[t-1,m],m]<-x[t,m]
}}

#hyper parameter
for (t in 2:T){
for (m in 1:M){
for (i in 1:I){
for (j in 1:J){
p[t,m,i,j]<-phi[t,m,i,j]/sum(phi[t,m,i,1:J])
}}}}

for (t in 2:T){
for (m in 1:M){
for (i in 1:I){
log(phi[t,m,i,5])<-0
for (j in 1:J-1){
log(phi[t,m,i,j])<-gamma1[t,j]+gamma2[t,i,j]+beta[i,j]*age[m]
}}}}

#define the component with multivariate normal
for (t in 2:T){
gamma1[t,1:J-1]~dmnorm(gamma1[t-1,1:J-1],W1[1:J-1,1:J-1])
}
gamma1[1,1:J-1]~dmnorm(mu.gamma1[1:J-1],W1[1:J-1,1:J-1])

for (t in 2:T){
for (i in 1:I){
gamma2[t,i,1:J-1]~dmnorm(gamma2[t-1,i,1:J-1],W2[1:J-1,1:J-1])
}}
for (i in 1:I){
gamma2[1,i,1:J-1]~dmnorm(mu.gamma2[i,1:J-1],W2[1:J-1,1:J-1])
}
for (i in 1:I){
beta[i,1:J-1]~dmnorm(mu.beta[i,1:J-1],W.beta[1:J-1,1:J-1])
}
}
```

```

W1[1:J-1,1:J-1]~dwish(R[1:J-1,1:J-1],4)
W2[1:J-1,1:J-1]~dwish(R[1:J-1,1:J-1],4)
W.beta[1:J-1,1:J-1]~dwish(R[1:J-1,1:J-1],4)

#loglikelihood
for (t in 2:T){
  for (m in 1:M){
    llike[t,m]<-log(p[t,m,x[t-1,m],x[t,m]])
  }
  for (t in 2:T){
    sumllike.t[t]<-sum(llike[t,1:M])
  }
}
#Deviance
D<--2*sum(sumllike.t[2:T])
}

```

References

- Anderson, T. W. and Goodman, L. A. (1957). Statistical inference about Markov chains, *The Annals of Mathematical Statistics*, **28**, 89–110.
- Bernardo, J. M. and Smith, F. M. (1994). *Bayesian Theory*, John Wiley & Sons, New York.
- Billingsley, P. (1961). Statistical methods in Markov chains, *The Annals of Mathematical Statistics*, **32**, 12–40.
- Duncan, G. and Lin, L. (1972). Inference for Markov chains having stochastic entry and exit, *Journal of the American Statistical Association*, **67**, 761–767.
- Lee, T. C., Judge, G. G. and Zellner, A. (1970). *Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*, North-Holland and Pub. Co., Amsterdam.
- Meshkani, M. R. and Billard, L. (1992). Empirical Bayes estimators for a finite Markov chain, *Biometrika*, **79**, 185–193.
- Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications, *Journal of American Statistical Association*, **78**, 47–65.
- Nhan, N. (1998). *Assessing Change Among Patients in Residential Treatment*, Technical Report, Graydon Manor Research Department, Virginia.
- Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996). *Bayesian Inference Using Gibbs Sampling Manual (version ii)*, MRC Biostatistics University, Cambridge University.
- Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit (with discussion), *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.
- Sung, M., Soyer, R. and Nhan, N. (2007). Bayesian analysis of non-homogenous Markov chains: Application to mental health data, *Statistics in Medicine*, **26**, 3000–3017.