A Bayesian Analysis of a Multiple Choice Test

by

Zhisui Luo

A Thesis

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Applied Statistics

by

April 24, 2013

APPROVED:

Professor Balgobin Nandram, Major Thesis Advisor

Abstract

In a multiple choice test, examinees gain points based on how many correct responses they got. However, in this traditional grading, it is assumed that questions in the test are replications of each other. We apply an item response theory model to estimate students' abilities characterized by item's feature in a midterm test. Our Bayesian logistic Item response theory model studies the relation between the probability of getting a correct response and the three parameters. One parameter measures the student's ability and the other two measure an item's difficulty and its discriminatory feature. In this model the ability and the discrimination parameters are not identifiable. To address this issue, we construct a hierarchical Bayesian model to nullify the effects of non-identifiability. A Gibbs sampler is used to make inference and to obtain posterior distributions of the three parameters. For a "nonparametric" approach, we implement the item response theory model using a Dirichlet process mixture model. This new approach enables us to grade and cluster students based on their "ability" automatically. Although Dirichlet process mixture model has very good clustering property, it suffers from expensive and complicated computations. A slice sampling algorithm has been proposed to accommodate this issue. We apply our methodology to a real dataset obtained on a multiple choice test from WPI's Applied Statistics I (Spring 2012) that illustrates how a student's ability relates to the observed scores.

Keywords: Item Response, Markov Chain Monte Carlo, Dirichlet Process Mixture

Acknowledgements

I would like to express my gratitude to my thesis advisor, Dr. Balgobin Nandram, who gave me such an interesting topic to work on; I have learned much more about Bayesian inference beyond the textbook and course. Thanks to Dr. Dilli Bhatta who helped me in revising my thesis. Also, I would like to thank Dr. Joseph D. Petrucelli who helped me a lot for the past three semesters.

Contents

1 Introduction

Item response theory (IRT) model is used to analyze the relationship between characteristics of exam items, examinees' abilities and test scores. In principle, the main idea of IRT model is to model the probability of getting correct response as a function of examinee's ability embedded with item features. Interestingly, those features are latent, and in most cases, we cannot make inference directly.

There are typically two different functions to formalize this relationship: normal function and logistic function. Of the two popular models, the latter one predominates in recent years, which is the one we will focus on. One of the practical and theoretical advantage of logistic function is that the cumulative distribution function has a closed form, which means the computation will be less expensive and less time-consuming.

The normal distribution was intuitively applied to describe the ability score and proportion of correct response, which is justified by Richardson (1936), Ferguson (1942) and Novick (1968) practically and theoretically. When the normal function is applied, discriminate and difficulty parameters can be easily expressed as location and scale parameters in cumulative normal distribution.

The logistic model has three extensions: the simplest and the most popular one is one-parameter logistic model, also known as the Rasch model. It contains only one variable-- the difficult parameter. In the one-parameter IRT model, items discriminate respondents in the same way for different probabilities of success for relative abilities (θ). The second type of logistic model, also known as two-parameter model, is considered both item discrimination (α) and difficulty (β) parameter. It is also the one we want to focus on in my thesis. A more complicated model is involved with another parameter, a guessing parameter in which an examinee gains extra points with successful guessing.

Traditional IRT model is divided into two families based on the kind of scored response: dichotomous and polytomous. In contrast to polytomous IRT models, which model the probability of selecting each response category, dichotomous IRT model deals with the probability of selecting correct response.

In binary logistic IRT model,

$$y_{ij} \mid \alpha_j, \beta_j, \theta_i \sim \text{Bernoulli}\left(\frac{e^{\alpha_j\theta_i-\beta_j}}{1+e^{\alpha_j\theta_i-\beta_j}}\right), \tag{1.1}$$

where $y_ij = 1$ if the i$^{\text{th}}$ student gets the j$^{\text{th}}$ problem correct and $y_{ij} = 0$ otherwise. $\{\alpha_j; j = 1, 2, \ldots, t\}$ is the discrimination parameter which illustrates the influence of students' ability on category propensity; $\{\beta_j; j = 1, 2, \ldots, t\}$ is the location parameter that reflects the difficulty of j^{th} item that does not depend on θ_i's; $\{\theta_i; i = 1, 2, \ldots, n\}$ is the individual ability for the i^{th} respondent. In principle, α is a scale parameter of abilities of individuals, therefore it is restricted to be positive. It is highly related to the item's difficulty value: under different levels of difficulties, discrimination parameters can be useful during different intervals of latent ability. Typically, its value varies from $1/2$ to 6.

The illustrated dataset is obtained from MA2611 Applied Statistics I-Test # 2. This course, which is an introductory statistics course, aims to let students gain a knowledge of basic statistical concepts, such as how to design and analyze experiments and sampling studies and how to analyze data in an appropriate way. The test we are going to analyze is the second test of this course. There are 17 questions, each question has four options and one of them is correct. There are 101 students in this class. After data screening and cleaning, the last question has two ambiguous answers, while all of students got the last but one problem correct, so we exclude them from further analysis.

Part of the dataset used in this paper is in Table 1. "1" represents that student got the correct answer and "0" means student does not get the correct answer.

Table 1. Sample dataset from MA2611 test

	Item										
	1	2	3	4	5	...	11	12	13	14	15
1	1	1	1	1	1	...	0	1	0	0	1
2	1	1	1	1	1	...	1	0	1	1	1
3	1	1	1	1	1	...	1	1	1	1	1
4	1	0	1	0	1	...	1	1	1	0	1
	...	...	...	...	...	...	...	...	...	...	...
98	0	0	0	1	1	...	1	1	1	0	1
99	1	1	1	1	0	...	1	1	1	1	1
100	1	0	1	1	1	...	0	0	1	1	0
101	1	1	1	1	1	...	1	0	1	0	1

We propose improved joint maximum likelihood estimation (JMLE) to specify the preliminary estimate for the MA2611 test data. A distinguishing characteristic of IRT model is that the α_j's and θ_i's can be subjectively scaled, that is estimation of α_j and θ_i in term "$\alpha_j\theta_i$"are determined by each other; to address this problem, Bolt *et al.* (2001) proposed an approach that restricts $\sum_{i=1}^{n}\theta_i = 0$. However, it used posterior parameter values twice, which caused "post-posterior" problem. We propose an improved JMLE algorithm to nullify such local identifiability.

The procedure of our algorithm is as follows:

Step 1 Roughly estimate θ_i, $i = 1, 2, \ldots, 101$ by

$$\hat{\theta}_i = \log(\frac{\bar{y}_{i\cdot} + 0.5}{1 - \bar{y}_{i\cdot} + 0.5}).$$

Step 2 Rescale θ as $\theta_1' = \hat{\theta}_1$, $\theta_i' = \hat{\theta}_i - \hat{\theta}_1, i = 2, 3, \cdots, 101.$

Step 3 Maximum likelihood estimate of α_j and β_j, $j = 1, 2, \ldots, 15$ using θ_i's are obtained using the following likelihood function:

$$L(\alpha_j, \beta_j \mid \boldsymbol{\theta}, \boldsymbol{y}) = \prod_{i=1}^{101} \frac{e^{(\alpha_j \hat{\theta}_i - \beta_j) y_{ij}}}{1 + e^{\alpha_j \hat{\theta}_i - \beta_j}}$$

Step 4 Estimate θ_i using $\hat{\alpha}_j$ and $\hat{\beta}_j$ with the restriction of $\theta_1 = 0$ by the same likelihood function.

Step 5 Go back to step 3 and step 4 until it converged.

Maximum likelihood procedure for estimating α_j's is L-BFGS-B proposed by Byrd et al. (1995) with respect to the positive restriction. It handles simple predictors with respect to box constraints which allows us to give bounds on α. The main procedure is to identify fixed and free variables at each step, and use only free ones to get higher accuracy based on L-BFGS, while we omit the detail here.

It turns out that the algorithm for the 16th item does not converge. It is probably due to the fact that all of the students except one got this question correct, which reveals that the question cannot discriminate students well and has relatively low difficulty. Therefore, we drop this question from further analysis. The values of estimation for α and β have been given in Table 2; for comparison, $\hat{p}$ is given to denote the proportion of correct response for each item.

Table 2. MLE estimates for α and β

	Item parameters				
	α	SE	β	SE	$\hat{p}$
1	0.432	0.558	-1.168	0.249	0.752
2	0.873	0.565	-0.127	0.218	0.505
3	3.458	1.034	-2.790	0.506	0.861
4	5.220	1.157	-1.989	0.411	0.693
5	6.839	1.656	-3.817	0.763	0.832
6	1.572	0.627	0.032	0.222	0.446
7	1.754	0.659	-0.927	0.255	0.663
8	3.270	0.827	-1.198	0.296	0.653
9	3.426	0.934	-2.178	0.405	0.802
10	3.299	1.056	-2.940	0.534	0.891
11	3.064	0.791	0.165	0.237	0.386
12	5.481	1.162	0.061	0.270	0.386
13	1.732	0.811	-2.035	0.362	0.842
14	2.373	0.700	-0.621	0.247	0.564
15	4.073	0.930	-1.088	0.301	0.604

From the preliminary study, the 4th and 5th problems have the most distinguishing ability, while the 11th question is the most difficult one among 15 questions, which is generally consistent with our knowledge about the data: one can check that fewest students hit the 11th problem.

An item characteristic curve (ICC) is used to visualize the relationship between ability level and the probability of correct response in a dichotomous IRT model. It shows the probability of a correct response as the function of student's ability. Sometimes, examinees with low abilities select the correct answer by chance, which could be characterized by a lower tail value of ICC asymptotic greater than zero; in which case, guessing correct response is more likely than expected. The three-parameter model is therefore applicable to deal with this situation.

As illustration, Figure 1 presents the probability of getting correct response corresponding to the extreme values of α and β with other parameters fixed. From the following ICC, we can better illustrate α and β: α reflects the determining power of an item: the higher the discrimination parameter, the greater increase in the proportion of correct response as a function of individual ability and the better the item is able to discriminate respondents for higher and lower levels; as for β, it points out that students are more likely to hit the question with lower level of difficulty than with higher one.

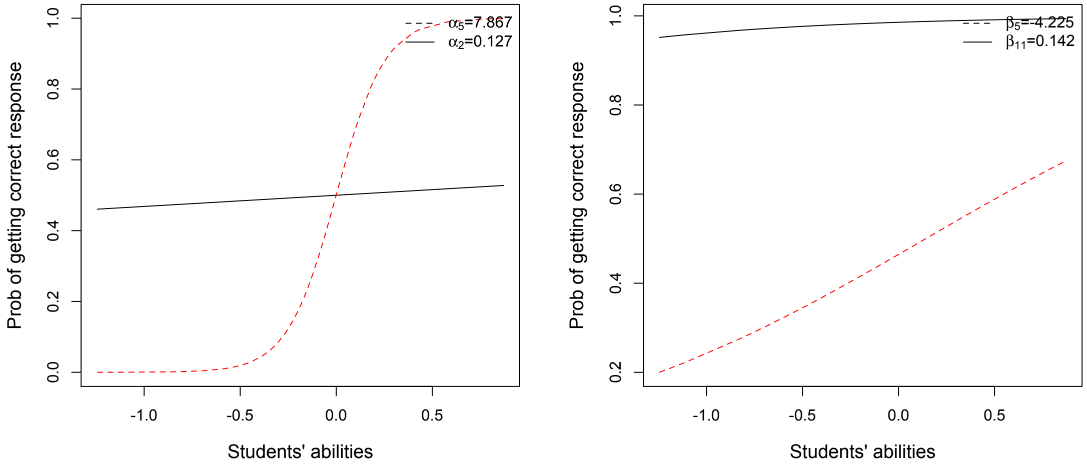

Figure 1. Item Characteristic Curve for each θ corresponding to fix α and β respectively

After preliminary analysis, the following section presents Gibbs Sampler applied to estimation of parameters of IRT model. Gibbs sampler is a relatively new Monte Carlo Markov Chain (MCMC) algorithm. It is used to obtain a sequence of observations from a complicated multivariate distribution when directly sample is difficult or even impossible.

Another problem we have worked on is how to cluster students based on their abilities obtained from IRT model. A straightforward way is to rank students in ascending (descending) order with no consideration of any variations. In comparison, we propose Dirichlet Process Mixture (DPM) model to cluster students automatically. The well-known DPM model is firstly introduced by Lo (1984).

In probability theory, a Dirichlet process, formally introduced by Thomas Ferguson (1973), is defined as a distribution over distributions. In general, Dirichlet process can be interpreted in four different perspectives: The Polya urn scheme, Chinese Restaurant (CRP), stick-breaking construction and formal definition. In this thesis, we focus on taking advantage of good cluster property of the interpretation of stick-breaking process.

One of approaches that are used to interpret Dirichlet process is stick-breaking construction, each component is drawn separately from base distribution.

Let us denote that $P \sim DP(\alpha_0, G)$, where G is a base function, and α_0 is a positive scaling parameter. The idea of DPM model is from this representation. A stick-processing representation is given by

$$P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}, \tag{1.2}$$

where $\theta_1, \theta_2, \theta_3, \ldots$ are independently and identically distributed from G and

$$\begin{cases} w_1 = \nu_1 & j = 1 \\ w_j = \nu_j \prod_{i<j}(1 - \nu_i) & j > 1 \end{cases} \tag{1.3}$$

and $\nu_j \sim$ Beta$(1, \alpha_0)$. δ_{θ_j} is a Dirac delta function, which denotes the measure with a point of 1 at θ_j.

The benefit of DPM model compared to traditional cluster algorithm is that the number of clusters can be determined by model and posterior results rather than pre-specified in modeling procedure; therefore, infinite mixture model has more favorable cluster property than other cluster methods. However, infinite number of components cause complicated algorithms and computations. Walker *et al.* (2011) introduced two latent variables to decide which components should be considered when processing Monte Carlo Markov Chain (MCMC) method.

In the rest of this paper we will discuss (1) parameters are estimated through Monte Carlo Markov chain method with ARS algorithm; (2) various model checking approaches will be discussed, including traceplots, autocorrelation, goodness-of-fit test and predictive test; (3) cluster students' abilities by DPM model with improved slice sampling method. The ideal output is to guide the instructor to grade students based on different levels of students' ability rather than only the proportion of correct responses.

2 Bayesian Estimation with MCMC Method

2.1 Specification of Model Parameter

Firstly, the joint likelihood function using model (1.1) is given by

$$f(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_i \prod_j \left[\frac{e^{\alpha_j \theta_i - \beta_j}}{1 + e^{\alpha_j \theta_i - \beta_j}} \right]^{y_{ij}} \cdot \left[\frac{1}{1 + e^{\alpha_j \theta_i - \beta_j}} \right]^{1 - y_{ij}}. \tag{2.1}$$

The following prior distributions are assumed for each IRT model parameter:

$$\begin{aligned}
&\alpha_j \overset{iid}{\sim} \text{Gamma}(\sigma_\alpha, \sigma_\alpha), j = 1, 2, \ldots, t, \\
&\beta_j \overset{iid}{\sim} N(\mu_\beta, \sigma_\beta^2), j = 1, 2, \ldots, t, \\
&\theta_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1, 2, \ldots, n, \\
&p(\mu, \sigma^2) \propto \frac{1}{(1 + \sigma^2)^2}, \\
&p(\sigma_\alpha) \propto \frac{1}{(1 + \sigma_\alpha)^2}.
\end{aligned}$$

Based on joint likelihood function and prior distributions, full posterior conditional distributions of all of the parameters can be determined by

$$\begin{aligned}
p(\alpha_j, \beta_j, \theta_i, \sigma_\alpha, \mu, \sigma^2 \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \alpha_j, \beta_j, \theta_i, \sigma_\alpha, \mu, \sigma^2) \cdot p(\alpha_j, \beta_j, \theta_i, \sigma_\alpha, \mu, \sigma^2) \\
&\propto p(\boldsymbol{y} \mid \alpha_j, \beta_j, \theta_i) \cdot p(\alpha_j | \sigma_\alpha) \cdot p(\beta_j) \cdot p(\theta_i \mid \mu, \sigma^2) \cdot p(\mu, \sigma^2) \cdot p(\sigma_\alpha)
\end{aligned}$$

$$= \prod_{i=1}^{n} \left[\frac{e^{\alpha_j \theta_i - \beta_j}}{1 + e^{\alpha_j \theta_i - \beta_j}} \right]^{y_{ij}} \cdot \left[\frac{1}{1 + e^{\alpha_j \theta_i - \beta_j}} \right]^{1 - y_{ij}} \cdot \text{Ga}(\sigma_\alpha, \sigma_\alpha) \cdot N(\mu_\beta, \sigma_\beta^2) \cdot \frac{1}{(1 + \sigma^2)^2} \cdot \frac{1}{(1 + \sigma_\alpha)^2}$$

We propose a MCMC technique, called Gibbs Sampler (Casella and George, 1992), to make inference of IRT model. The Gibbs Sampler generates a random sample from $\{x_1, x_2, \ldots, x_n\}$ from the joint distribution $p(x_1, x_2, \ldots, x_n)$ as follows:

1. Let $\{x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\}$ be starting values.

2. Draw $x_i^{(j)}$ from condition distribution $p(x_i | x_1^{(j)}, \ldots, x_{i-1}^{(j)}, x_{i+1}^{(j-1)}, \ldots, x_n^{(j-1)})$. That is, draw each variable from the conditional distribution the most recent values and updating the variable with its new value after it has been sampled.

After a large number, B, of iterations, we obtain $\{x_1^{(B)}, x_2^{(B)}, \ldots, x_n^{(B)}\}$.

It is noted that we assume α_j follows a gamma distribution, which is different from "normal distribution" in popular researches. The reason is that it is ensured the mean of α_j always equal to 1, so the inference will not be affected by its neighbor θ_i.

The Gibbs sampling method requires full conditional distributions of all parameters and hyperparameters to generate random samples from this posterior densities. In the following, full conditional posterior densities can be obtained as follows:

1. Sample item latent parameters for all categories. It is assumed that items are conditionally independent. The parameter $\{\alpha_j; j = 1, 2, \ldots, t\}$ and $\{\beta_j; j = 1, 2, \ldots, t\}$ have the following conditional distributions

$$\begin{aligned}
\pi(\alpha_j \mid \boldsymbol{y}, \beta_j, \boldsymbol{\theta}, \sigma_\alpha, \mu, \sigma^2) &\propto \prod_{i=1}^{n} \left[\frac{e^{\alpha_j\theta_i - \beta_j}}{1 + e^{\alpha_j\theta_i - \beta_j}}\right]^{y_{ij}} \cdot \left[\frac{1}{1 + e^{\alpha_j\theta_i - \beta_j}}\right]^{1-y_{ij}} \cdot \text{Ga}(\sigma_\alpha, \sigma_\alpha), \\
\pi(\beta_j \mid \boldsymbol{y}, \alpha_j, \boldsymbol{\theta}, \sigma_\alpha, \mu, \sigma^2) &\propto \prod_{i=1}^{n} \left[\frac{e^{\alpha_j\theta_i - \beta_j}}{1 + e^{\alpha_j\theta_i - \beta_j}}\right]^{y_{ij}} \cdot \left[\frac{1}{1 + e^{\alpha_j\theta_i - \alpha_j}}\right]^{1-y_{ij}} \cdot N(\mu_\beta, \sigma_\beta^2).
\end{aligned}$$

2. Sample latent ability parameter for all of students. Similarly, we assume all $\{\theta_i; i = 1, 2, \ldots, n\}$ are conditionally independent. The latent ability has the

following conditional distributions

$$\pi(\theta_i \mid \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma_\alpha, \mu, \sigma^2) \quad \propto \quad \left\{ \prod_{j=1}^{t} \left[\frac{e^{\alpha_j \theta_i - \beta_j}}{1 + e^{\alpha_j \theta_i - \beta_j}} \right]^{y_{i\cdot}} \cdot \left[\frac{1}{1 + e^{\alpha_j \theta_i - \beta_j}} \right]^{1-y_{i\cdot}} \right\} \cdot N(\mu, \sigma^2).$$

3. Sample class ability means and variances μ and σ^2 for each class. It is assumed that the ability parameters are independent for all other variables except the θ_i,

$$\mu \mid y_{ij}, \boldsymbol{\theta}, \sigma^2 \sim N(\bar{\theta}, \frac{\sigma^2}{n}),$$
$$\sigma^2 \mid \boldsymbol{\theta} \sim Inv - \chi^2(n-1, s^2),$$
$$p(\mu|\boldsymbol{\alpha}) \propto \prod_{j=1}^{15} \left[\alpha_j^{\alpha-1} \cdot e^{-\alpha_j} \cdot \frac{1}{(1+\alpha_j)^2} \right].$$

It can be seen that sampling from μ and σ^2 are easy to be conducted based on their conditional distributions, whereas we only know up to a normalizing constant with respect to α_j, β_j and θ_i. We noticed that those distributions are log-concave, which allows us to sample random deviate of these parameter using Adaptive Rejection Sampling (ARS). Adaptive Rejection Sampling is a sample technique from any univariate log-concave density distribution. The log-concavity proof is provided in Appendix 1. ARS algorithm could be realized by WinBUGS project.

2.2 Illustrative Example

We use MA2611 test data as illustration. The Markov chains for each parameter runs for $15,000$ iterations with the first 1000 past burn-in iterations. The rest $14,000$ iterations were used to compute parameter estimates. Table 3 and 4 show estimates of α, β and hyperparameters of θ. The last column is the actual proportions of correct responses for all items. The starting values are chosen from JMLE estimation in previous section.

Table 3. Item category estimates for α and β

	Item parameters estimates				
	$\hat{\alpha}$	SE	$\hat{\beta}$	SE	$\hat{p}$
1	0.312	0.245	-0.799	0.286	0.752
2	0.231	0.191	0.163	0.247	0.505
3	1.682	0.715	-0.785	0.461	0.861
4	1.616	0.637	0.271	0.449	0.693
5	2.703	1.132	-0.250	0.554	0.832
6	0.303	0.230	0.445	0.271	0.446
7	0.507	0.326	-0.220	0.320	0.663
8	1.158	0.508	0.255	0.411	0.653
9	1.169	0.518	-0.509	0.403	0.802
10	1.430	0.645	-1.032	0.446	0.891
11	0.771	0.391	1.060	0.373	0.386
12	0.923	0.436	1.228	0.399	0.386
13	0.810	0.444	-0.997	0.379	0.842
14	0.691	0.369	0.282	0.348	0.564
15	1.305	0.521	0.572	0.407	0.604

Table 4. Hyperparameter estimation for θ

parameter	mean	SE
μ	0.866	0.240
σ	1.901	0.456

In comparison with preliminary analysis, values of discrimination parameters α are actually pretty similar; in reality, this result makes more sense: respondents in our analysis come from the same class and learn from the same instructor, which means technically, test questions are likely to treat students in a more even way. For β, all of questions are kind of below "medium difficulty", while most people got 10th and 13th questions correct. Figure 2 illustrates how students' abilities vary with respect to the probability of getting correct response. It shows that students who have got the same proportion of correct responses have different levels of abilities if considered item's feature.

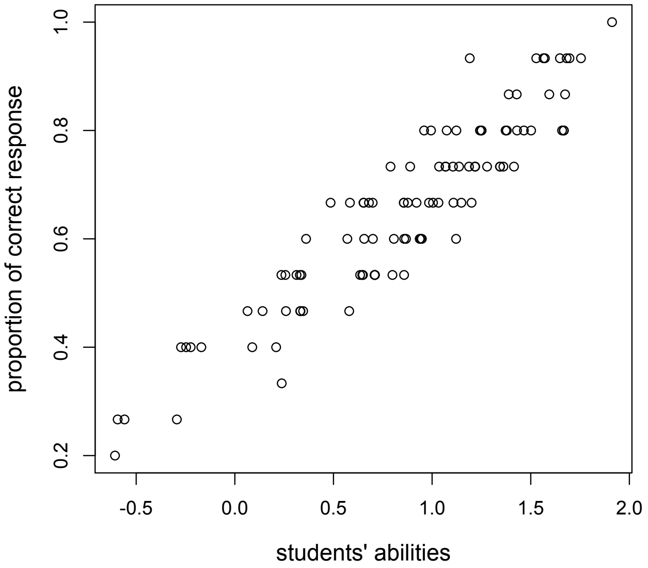

Figure 2. Relationship between estimation of θ and the proportion of correct response for students

2.3 Model Diagnosis

The nature of MCMC algorithm has two main issues: firstly, some initial iterations are to be discarded because the algorithm needs some iterations to reach convergence; secondly, MCMC samples need a long enough time to be independent from the past in the sequence, which results in difficulty of estimating the error variance of the MCMC estimate. There are two ways to assess convergence: trace plots and autocorrelations. In order to test how well our model fit, we propose a goodness-of-fit test using a family of divergence statistic. At last, Bayesian cross-validation is used to examine the predictive ability of our model in view of both items and individuals.

The most popular way of detecting the nonstationary behavior of the chain is traceplot. It illustrates the history of chain simulated by ARS with respect to all item parameters and individual abilities. Of note, we used hyperparameters μ and σ^2 as examples.

Figure 3 shows an example of traceplots for ability precision parameters. It provides a strong evidence that both μ and σ have gone through a transient state.

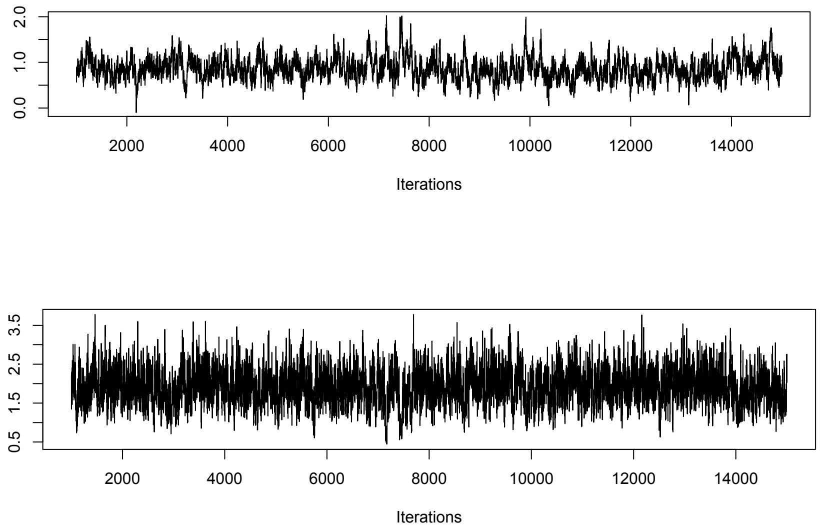

Figure 3. MCMC history of μ and σ^2

The autocorrelation is defined as:

Let $U^{(1)}, U^{(2)}, U^{(3)}, \cdots, U^{(N)}$ be the N iterations for variable U after initial convergence. The autocorrelation coefficient of the sequence $\{U^{(i)}\}$ at lag k is

$$r_k = \sum_{i=1}^{N-k} \frac{(U^{(i)} - \bar{U})(U^{(i+k)} - \bar{U})}{\sum_{i=1}^{N}(U^{(i)} - \bar{U})^2} \tag{2.2}$$

where $\bar{U} = N^{-1}\sum_{i=1}^{N} U^{(i)}$, and its asymptomatic standard error is $ste_k = \left\{ \frac{N-k}{N(N+2)} \right\}^{1/2}$.

Figure 4 reveals the autocorrelation plots for μ and σ^2. The correlation of σ^2 washed out quickly after several lags, which indicates a relatively low correlated samples; whereas μ has a relatively higher correlation seen from the plot. It is noted that those plots for α, β and θ have also been examined, although we do not present here.

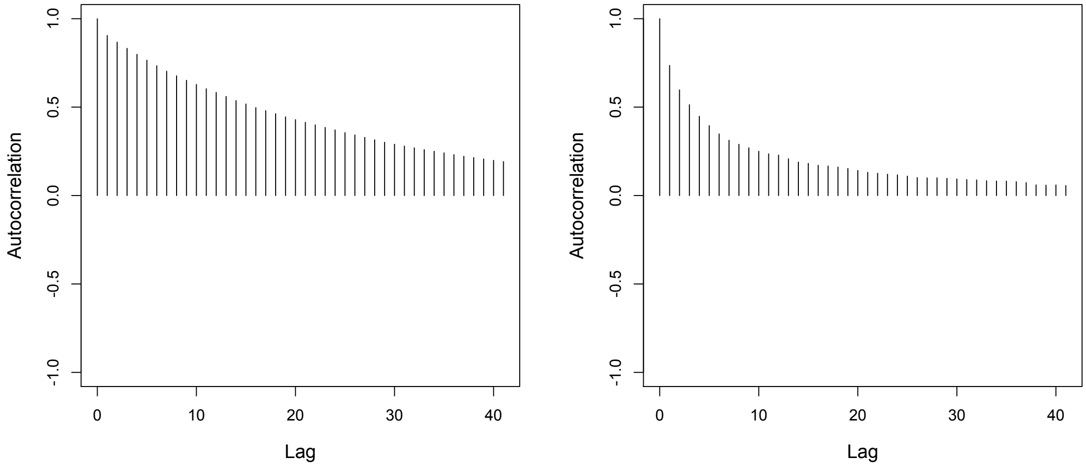

Figure 4. Autocorrelations for μ and σ^2

The goodness-of-fit test describes how well the model fits with a summary measure of discrepancy between observed values and the expected values. The most widely used statistic in goodness-of-fit test is Pearson's χ^2. In order to check how well our model fit, we want to check the goodness-of-fit using Cressie-Read (1984) statistic. They investigated a family of power divergence statistic testing the goodness-of-fit. Here, we used a family of statistics with different λ's.

For testing how well the model fits for students' abilities, we define

$$T_\lambda = \frac{2}{\lambda(1+\lambda)} \sum_{j=1}^{15} y_{ij} \left[\left(\frac{y_{ij}}{n_j p_j} \right)^{\lambda} - 1 \right], \qquad (2.3)$$

where y_{ij} is the observed number of correct response and $n_j p_{\boldsymbol{j}}$ is the expected number, which is approximated by $E(T = \sum_i y_{ij})$. The special cases are likelihood ratio statistic ($\lambda = -1$) and Pearson's χ^2 statistic ($\lambda = -2$).

Under the Bayesian context, lack of fit data will be measured by posterior predictive p-value, that is

$$p_B = Pr\left\{T_\lambda\left({y_{ij}}^{rep}, \boldsymbol{p}\right) > T_\lambda\left({y_{ij}}^{obs}, \boldsymbol{p}\right) \mid y_{ij}\right\}. \tag{2.4}$$

A model may be lack-of-fit when the posterior p-value is close to 0 or 1. It indicates that the model cannot provide replications of the data. A similar statistic can be used for testing the items.

Figure 5 shows the posterior predictive p-value versus λ corresponding to items and individuals. Most of p-values are around .5 for students, which is a strong evidence that our model fits quite well; while for items, most of p-values shows favorbale results for $\lambda \in (-1, 1)$.

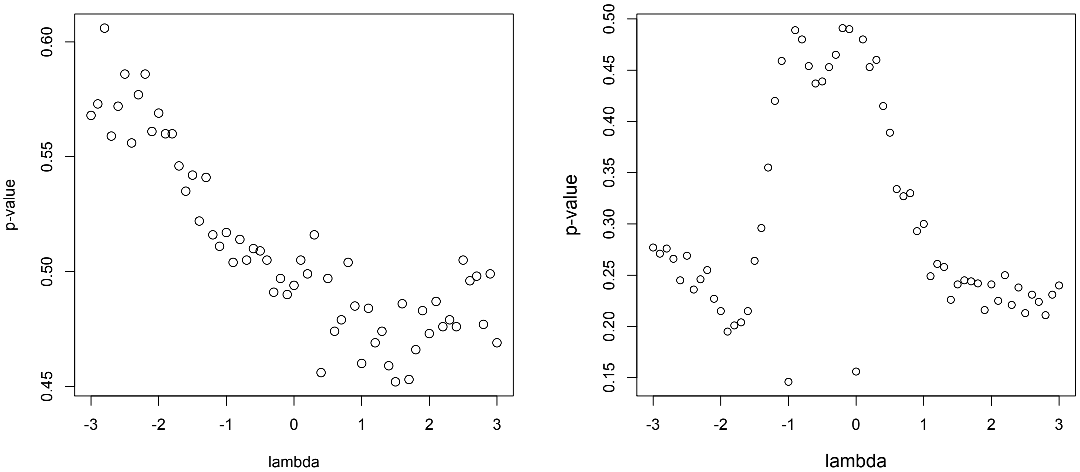

Figure 5. Goodness-of-fit test for students and items

At last, we focus on assessing the predictive ability of our model. For predicting items, we denote $\boldsymbol{Y}_j$ as the (101×1) vector for each item, and $\boldsymbol{Y}_{(j)}$ denote the vector with the j^{th} item deleted.

We use $\boldsymbol{Y}_{(j)}$ and predict $\boldsymbol{Y}_j$ using the fitted model. If it fits well, it should predict $\boldsymbol{Y}_j$ very well. Here, we applied a Normal Probability Plot of residuals to show the goodness of fit.

Let's denote that $T_j = \sum_i y_{ij}$, then the standardized deleted residuals are

$$Dres_j = \frac{T_j - E(T_j \mid \boldsymbol{Y}_{(j)})}{\sqrt{var(T_j \mid \boldsymbol{Y}_{(j)})}}, \tag{2.5}$$

where

$$E(T_j \mid \boldsymbol{Y}_{(j)}) = \sum_{k=1}^{M} w_j^{(k)} E(T_j \mid \boldsymbol{p}^{(k)}) \tag{2.6}$$

and

$$w_r^{(k)} = \frac{\left[f(T_j \mid \boldsymbol{p}^{(k)})\right]^{-1}}{\sum_{k=1}^{M} \left[f(T_j \mid \boldsymbol{p}^{(k)})\right]^{-1}} \tag{2.7}$$

with $\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}, \cdots, \boldsymbol{p}^{(M)}$ a random sample from $\pi(\boldsymbol{\theta} \mid y_{ij})$ A new approach to calculate weight is in Appendix 2. Calculation of deleted student's residual is similar.

The residual plots predicting all deleted items and students are shown in the Figure 6.

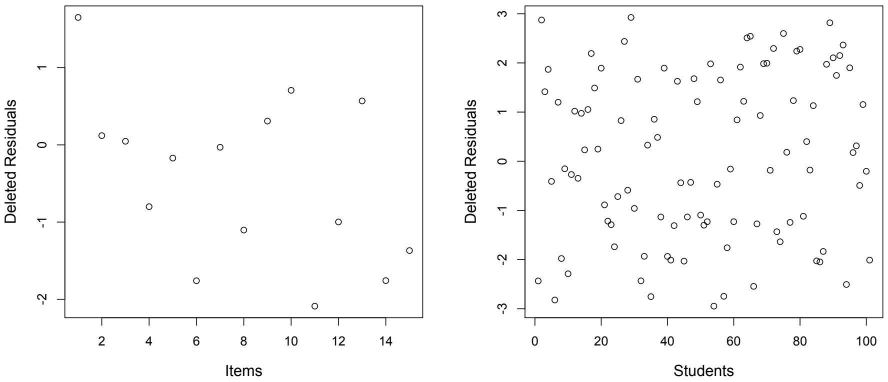

Figure 6. Deleted Residual Plot for all items

3 Dirichlet Process Mixture Model

3.1 Introduction

Dirichlet process mixture (DPM) model is firstly introduced by Lo(1984) with the Gaussian kernel

$$f_P(y) = \int N(y;\theta)dP(\theta), \tag{3.1}$$

where $P \sim DP(\alpha_0, G)$ and $\theta = (\mu, \sigma^2)$. With inspiration of stick-processing representation, we can write an infinite-dimensional mixture model, whose each component is drawn separately from base distribution, that is:

$$f_{\theta,w}(\boldsymbol{y}) = \sum_{j=1}^{\infty} w_j \cdot p(\boldsymbol{y};\theta_j|\boldsymbol{\alpha},\boldsymbol{\beta}), \tag{3.2}$$

where

$$p(\boldsymbol{y};\theta_j|\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{m=1}^{t} \left[\frac{e^{\alpha_m\theta_j-\beta_m}}{1+e^{\alpha_m\theta_j-\beta_m}}\right]^{y_{\cdot m}} \cdot \left[\frac{1}{1+e^{\alpha_m\theta_j-\beta_m}}\right]^{1-y_{\cdot m}}. \tag{3.3}$$

Our purpose is to implement Gibbs sampler in this joint likelihood function; however, it is quite difficult to sample infinite number of θ_j's to proceed algorithm.

Walker *et al.* (2011) proposed a new sampler method for sampling the DPM model. This approach introduces two latent variables which make finite number of mixtures, which tried to avoid such difficulties. The key idea of slice sampler is to introduce a latent variable u that help to sample from a finite number of θ_j's. After given u, the number of partitions becomes finite:

$$f_{\theta,w}(\boldsymbol{y}, u) = \sum_{j=1}^{\infty} \mathbf{1}(u < w_j) \cdot p(\boldsymbol{y}; \theta_j|\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{3.4}$$

When given u, the indices are reduced to $A_u = \{j : w_j > u\}$. After making a finite sum, a further latent variable, d, will be introduced. It indicates which component attribute to the density function, with which trick we can get rid of summation sign:

$$f_{\theta,w}(\boldsymbol{y}, u, d) = \mathbf{1}(u < w_d) \cdot p(\boldsymbol{y}; \theta_j|\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{3.5}$$

where w is defined in (1.3).

Although this form of distribution could be easily handled by Gibbs Sampler, there are some limitations that will cost extra works. When proceeding Gibbs sampling, updating u will cause the change of the set $\cup_{i=1}^{n} A(u_i)$ and consequently lead to more simulation of w's.

To overcome this problem, we can fix w's in the generation of u. Therefore, a more general form had been proposed:

$$f_{\nu,\theta}(\boldsymbol{y}, u, d) = \xi_d^{-1} \mathbf{1}(u < \xi_d) w_d p(\boldsymbol{y}; \theta|\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{3.6}$$

where $\xi_1, \xi_2, \xi_3, \ldots$ is any positive sequence. The choice of this sequence is another issue that we will not discuss here. Here, we consider the sequence as $\xi_j = (1-k)k^{j-1}$ where $k = 0.5$. The joint likelihood function is defined by

$$L_{w,\theta}(\{y_i, u_i, d_i = k_i\}_{i=1}^{101}) = \prod_{i=1}^{n} f_{\nu,\theta}(\boldsymbol{y_i}, u_i, d_i) = \prod_{i=1}^{n} \left\{\xi_{d_i}^{-1} \mathbf{1}(u_i < \xi_{d_i}) w_{d_i} p(\boldsymbol{y_i}; \theta_j|\boldsymbol{\alpha}, \boldsymbol{\beta})\right\}. \tag{3.7}$$

The prior distributions are defined

$$\theta_j \sim N(\mu, \sigma^2),$$
$$p(\mu, \sigma^2) \propto \frac{1}{(1+\sigma^2)^2},$$
$$\nu_j \sim \text{Beta}(1, \alpha_0).$$

3.2 Sampling Algorithm

In this section, we are going to implement Gibbs Sampling for the proposed density distribution. The variables that need to be sampled are $\{(\theta_j, \nu_j), j = 1, 2, \ldots; (d_i, u_i), i = 1, 2, 3, \ldots, n; (\mu, \sigma^2)\}$.

1. We begin with sample the latent variable u_i which is also the simplest one. The condition posterior distribution is

   $$\pi(u_i|\cdots) \propto \mathbf{1}(0 < u_i < \xi_{d_i}).$$

2. Then, we sample the weight of mixture model ν_j. Based on the joint likelihood distribution (3.8), the conditional distribution of ν_j is given by

   $$\pi(\nu_j|\cdots) \propto \text{Beta}(a_j, b_j),$$

   where $a_j = 1 + \sum_{i=1}^n \mathbf{1}(d_i = j)$ and $b_j = \alpha_0 + \sum_{i=1}^n \mathbf{1}(d_i > j)$, and hence we can calculate w_j's.

3. We will sample individual abilities θ_j in this step. The posterior conditional

distribution is as follows:

$$f(\theta_j|\cdots) \propto N(\mu,\sigma^2)\prod_{d_i=j} p(\boldsymbol{y_i};\theta_j|\boldsymbol{\alpha},\boldsymbol{\beta}),$$

when there is no $\{k_i = j\}$, then

$$\theta_j|\cdots \sim N(\mu,\sigma^2).$$

4. Then we will sample the indicator variables d_i. It is given by

$$p(d_i = k|\cdots) \propto \mathbf{1}(k : \xi_k > u_i)w_k/\xi_k N(y_i;\theta_k).$$

5. At last, the posterior conditional distribution for α_0 is

$$\pi(\alpha_0|d,\cdots) \propto \alpha_0^d \Gamma(\alpha_0)\pi(\alpha_0)/\Gamma(\alpha_0+n),$$

where d is the number of distinct k_i's, that is the number of clusters. We will present it will be a nice way to sample from the posterior distribution when prior distribution of α_0 is a gamma distribution. Suppose $\alpha_0 \sim$ Gamma(a,b), then we can deduce

$$\begin{aligned}\pi(\alpha_0|d,\cdots) &\propto \pi(\alpha_0)\alpha_0^{d-1}(\alpha_0+n)\mathrm{B}(\alpha_0+1,n)\\ &\propto \pi(\alpha_0)\alpha_0^{d-1}(\alpha_0+n)\int_0^1 x^{\alpha_0}(1-x)^{n-1}dx,\end{aligned}$$

$\pi(\alpha_0|d,\cdots)$ can be the marginal distribution of the following distribution:

$$\pi(\alpha_0|d,\eta,\cdots) \propto \pi(\alpha_0){\alpha_0}^{d-1}(\alpha_0+n)\eta^{\alpha_0}(1-\eta)^{n-1},$$

where $\alpha > 0$ and $0 < \eta < 1$. After simple algebra, the posterior distribution of α_0 reduces to the mixture of two gamma distributions,

$$\pi(\alpha_0|d, \eta, \cdots)\pi_\eta G(a+d, b-\text{log}(\eta)) + (1-\pi_\eta)G(a+d-1, b-\text{log}(\eta)),$$

where $\pi_\eta/(1-\pi_\eta) = (a+d-1)/[n(b-\text{log}(\eta))]$. Second,

$$\pi(\eta|\alpha_0, d) \propto \eta^{\alpha_0}(1-\eta)^{n-1},$$

that is, posterior distribution of η follows a beta distribution with mean $(\alpha_0 + 1)/(\alpha_0 + n + 1)$.

After obtaining d and α_0 during each iteration, draw η from beta distribution and then update α_0 from the mixture gamma distributions using ARS algorithm (log-concavity proved). Since μ and σ^2 only depend on θ, their posterior distributions are as the same as in section 2.3.

To succeed proceeding algorithm, we need to sample enough θ_j's. The principle to find required set of k is $k = \{1, 2, \ldots, N\}$, $N = \text{max}_i\{N_i\}$, where N is the largest integer l for which $\xi_l > u_i$.

3.3 Application of DPM Model into MA2611 Test Data

In this section, of interest is to cluster students based on different ability levels. To initialize d_i's, the students' abilities were split into 10 equal clusters according to ascending order. Student's ability parameter θ_i with μ and σ^2 will be the only parameter when directly retrieved values of discrimination and difficulty parameter in Section 2.2. The Gibbs Sampler ran for 5000 iterations with the first 500 burn-in samples. Figure 7 shows the running average of number clusters, it is clear that 5000 iterations is good enough to reach the stable condition.

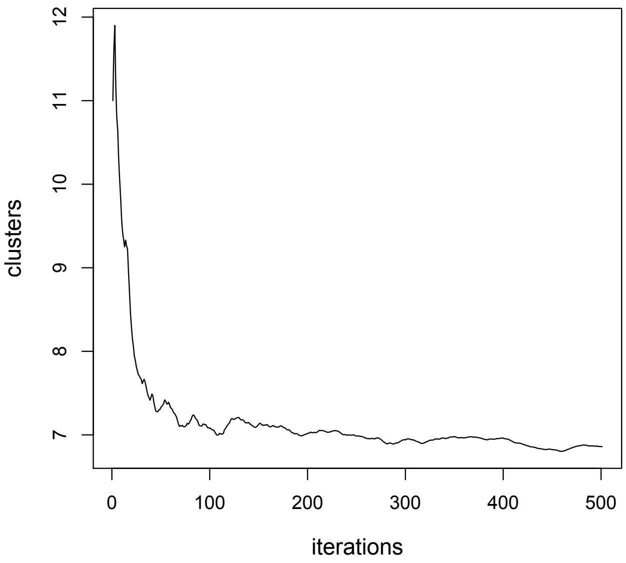

Figure 7. Running average for the number of clusters

Figure 8 shows the frequency of latent number of clusters, that is the distinct number of d_i's denoted by K, and the history of Gibbs Sampler; it can be seen that the potential clusters are most frequently appeared as 5, 6 and 7.

We will consider to cluster students based on the specific K. Although K occurs at pretty high values, d_i will only appear much lesser times due to the randomly selection at some iterations through the whole process; to accommodate this situation, my choice is to choose the most frequent d_i's as i^{th} group indicator for all of iterations under specific condition of K. The following three plots presents the group indicators chosen by the above criteria versus students' abilities under the most frequent number of clusters $K = 5, 6, 7$ in Figure 9.

Table 5 shows part of group indicators when distinct groups are 4, 5, 6 and 7.It is pleasant to see that most items share the same group indicators among four different

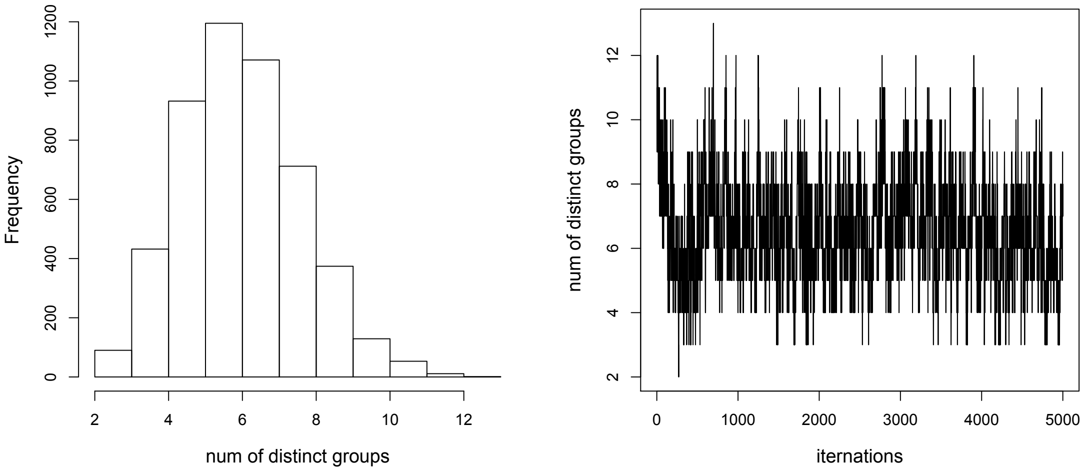

Figure 8. Frequency of number of clusters

clusters. As result, the actual number of clusters of students is less than 4 under most iterations although the potential number is higher.

It is also noted that most students were assigned into the same group (i.e. 78 out 101 students in the same group when $K = 6$), which means most students in this class have similar ability for this test.

3.4 Model Diagnosis

We applied similar model diagnosis procedures in Section 2.3. We provided such diagnosis for all parameters but only use two of them as illustration. Figure 10 and 11 show traceplots and autocorrelation plots for hyperparameters of θ with respect to the results in Section 3.3.

Figure 12 shows the posterior predictive p-value versus λ corresponding to items and individuals.

The residual plots for predicting deleted items and students are shown in the Fig-

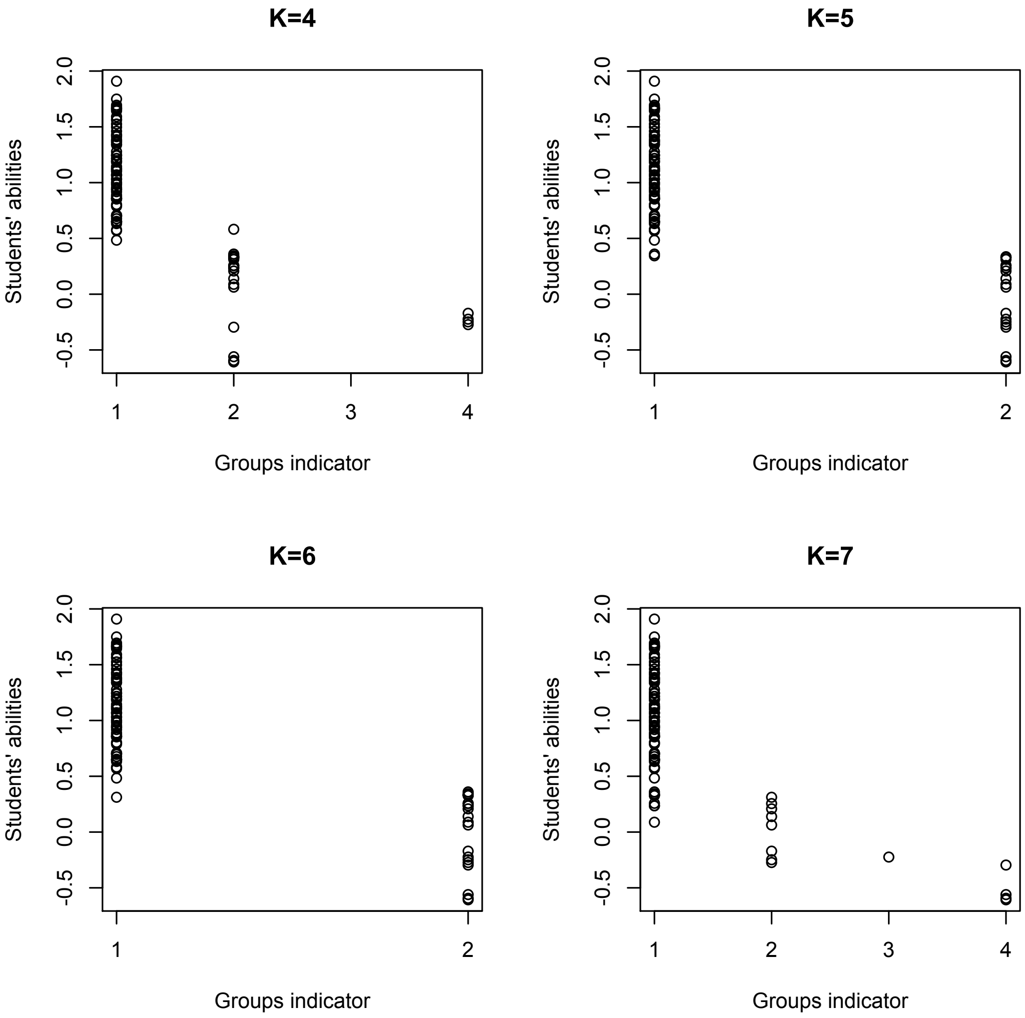

Figure 9. Frequency of number of clusters

Table 5. Groups Indicators under Different Numbers of Clusters

	Number of Distinct Groups			
	$K=4$	$K=5$	$K=6$	$K=7$
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	2	2	2	1
8	4	2	2	2
...	...	...	...	...
100	1	1	1	1
101	1	1	1	1

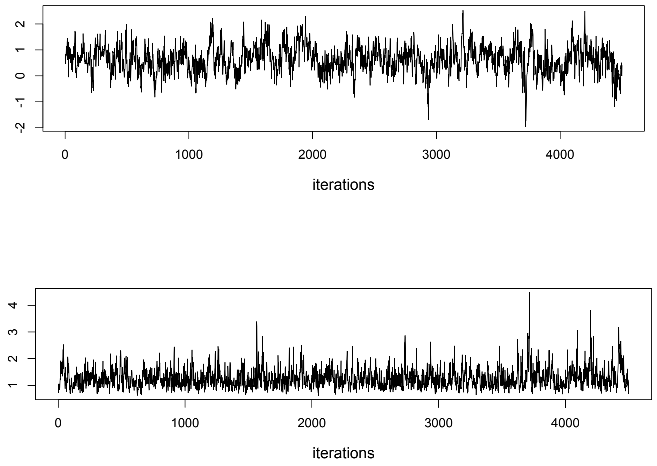

Figure 10. Traceplots for μ and θ

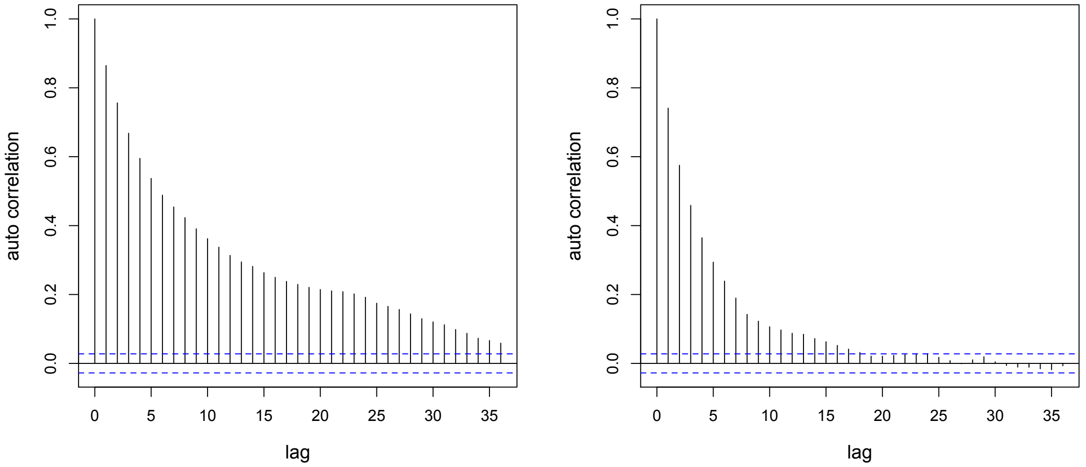

Figure 11. Autocorrelation Plots for μ and θ

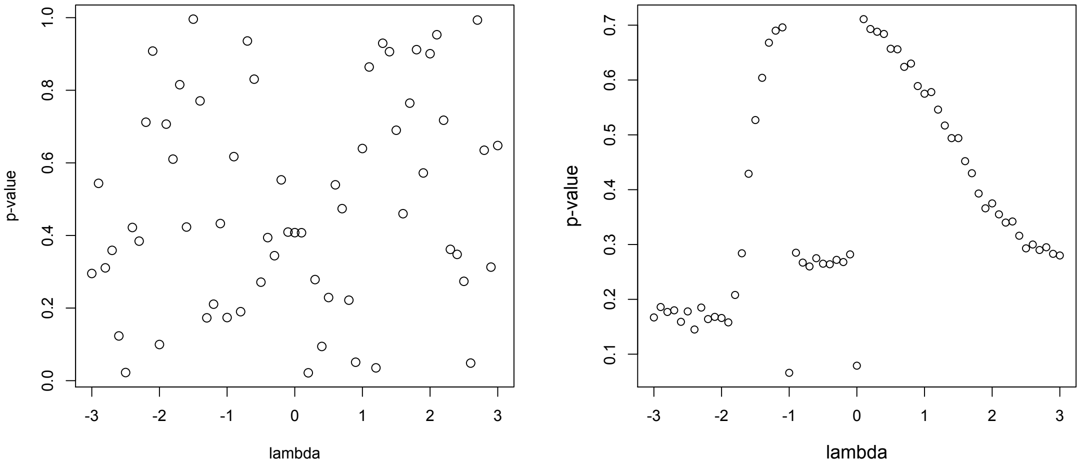

Figure 12. Goodness-of-fit test for fitting items and students

ure 13. Of note, the residual plots show better random measurement errors than those in Section 2.5.3.

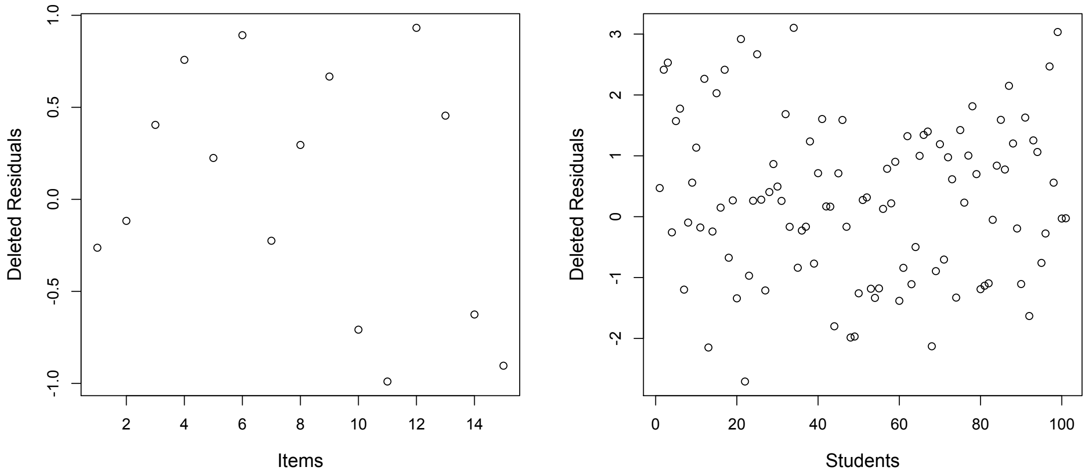

Figure 13. Deleted Residuals Plot for items and students

4 Conclusion

We have discussed binary Item Response Theory model and Dirichlet process mixture (DPM) model. We have applied both models to MA2611 test # 2, Spring 2012, and as an illustration, several model diagnosis procedures followed. The primary objective is to discuss two-parameter IRT model of dichotomous response and cluster students automatically with DPM model. It is highly recommended to grade students in this way, since both student's ability and item's features are taken into consideration, instead of just the proportion of correct responses. Firstly, we have provided a Markov chain Monte Carlo (MCMC) method with Adaptive Rejection Sampling (ARS) to estimate parameters. Using this method, we found that students' abilities were consistent with the number of correct responses they got. Two improved model checking procedures have been proposed. Except for traditional traceplots and autocorrelation, a family of divergence statistics, Cressie-Read statistic, was used to test the goodness-of-fit of our model. The evidence of goodness-of-fit therefore becomes stronger.

DPM model has good property of clustering, since we do not have to decide specific number of clusters in advance compared to a finite mixture model. However, infinite discrete components from a random distribution can cause expensive computation. To overcome this issue, slice sampling algorithm introduces two latent variables to determine which components are required to be sampled, which results in a pretty simple easy and simple formation, see Kalli *et al.* (2011).

By building similar full conditional distributions, we implemented students' abilities obtained from IRT model into DPM model. Potential number of clusters were then obtained that mostly occurred with sizes 5, 6 and 7. The final decision is based on the expertise of the instructors.

Future work will focus on how to grade students more precisely. Specifically, using the Gibbs sampler, "who belongs to which group" still remains an issue due to the nature of MCMC although we can obtain potential number of clusters. In which case, students can be graded with uncertainty; that is, student A can be possibly graded with student B who, however has lower ability. To accommodate this flexibility, an instructor may consider combining our result with in-class performance and subjective impression when grading.

A Appendix

A.1 Log-concavity of Full Conditional Distributions

There is a prerequisite when performing Gibbs Sampling using ARS algorithm: it can be only applied to conditional distributions that are log-concave. A density is called log-concave if its second derivative is non-negative everywhere. Here, we prove conditional distribution of α_j and it is similar for β_j and θ_i.

$$\begin{aligned}\pi(\alpha_j \mid y_{ij}, \beta_j, \theta_i, \mu, \sigma^2) \quad &\propto \quad \prod_{i=1}^{n}\left[\frac{e^{\alpha_j\theta_i-\beta_j}}{1+e^{\alpha_j\theta_i-\beta_j}}\right]^{y_{ij}} \cdot \left[\frac{1}{1+e^{\alpha_j\theta_i-\beta_j}}\right]^{1-y_{ij}} \cdot \frac{1}{\sqrt{2\pi\sigma_\alpha^2}}e^{-\dfrac{(\alpha_j-\sigma_\alpha)^2}{2\sigma_\alpha^2}} \\ &\propto \quad \prod_{i=1}^{n}\frac{e^{(\alpha_j\theta_i-\beta_j)y_{ij}}}{1+e^{\alpha_j\theta_i-\beta_j}} \cdot \frac{1}{\sqrt{2\pi\sigma_\alpha^2}}e^{-\dfrac{(\alpha_j-\sigma_\alpha)^2}{2\sigma_\alpha^2}}.\end{aligned}$$

Taking logarithm transformation on both sides, and second derivatives of the log-likelihood,

$$\begin{aligned}&\ln\pi(\alpha_j) = \sum_{i=1}^{n} y_{ij}(\alpha_j\theta_i-\beta_j) - \sum_{i=1}^{n}\ln(1+e^{\alpha_j\theta_i-\beta_j}) - \frac{(\alpha_j-\sigma_\alpha)^2}{2\sigma_\alpha^2}, \\ &\frac{\mathrm{d}}{\mathrm{d}\alpha_j}\ln\pi(\alpha_j) = \sum_{i=1}^{n}\alpha_j y_{ij} - \sum_{i=1}^{n}\frac{\theta_i e^{\alpha_j\theta_i-\beta_j}}{1+e^{\alpha_j\theta_i-\beta_j}} - \frac{(\alpha_j-\sigma_\alpha)}{\sigma_\alpha^2}, \\ &\frac{\mathrm{d}^2}{\mathrm{d}{\alpha_j}^2}\ln\pi(\alpha_j) = -\frac{1}{\sigma^2} - \sum_{i=1}^{n}\frac{\theta_i^2 e^{\alpha_j\theta_i-\beta_j}}{(1+e^{\alpha_j\theta_i-\beta_j})^2} < 0.\end{aligned}$$

The log-concavity has been proved.

A.2 Approach to Calculate Exact Weights

The key point to calculate the weight is to give out $f(y_r \mid \boldsymbol{p}^{(k)})$. Firstly, Monte Carlo simulation was attempted to estimate the density approximately; however, sometimes the random samples drawn from density distributions failed to match the observed data, and hence yielded the estimated density to be 0.

Then an exact way to calculate the density distribution has been proposed. Let's denote $f(T_j \mid \boldsymbol{p}^{(k)}) = \Pr(T_j = t)$,

$$\begin{aligned}\Pr(T_i = t) &= \Pr(\boldsymbol{Y_1} = t) \cup \Pr(\boldsymbol{Y_2} = t) \cup \cdots \cup \Pr(\boldsymbol{Y}_{\binom{n}{t}} = t)\\ &= \sum_{y_{ij}:T_j=t} I\left(T_i = t\right) P(\boldsymbol{Y}).\end{aligned}$$

Our purpose is to get rid of indicator function, then

$$\begin{aligned}&= \textstyle\sum_{y_{ij}:T_j=t} I\left(T_j = t\right) \frac{P(\boldsymbol{Y})}{P_e(\boldsymbol{Y})} \cdot P_e(\boldsymbol{Y}) = \sum_{y_{ij}:T_j=t} \cdot \frac{P(\boldsymbol{Y})}{P_e(\boldsymbol{Y})} \cdot P_e(\boldsymbol{Y}) \cdot I\left(T_j = t\right)\\ &= \textstyle\sum_{y_{ij}:T_j=t} \left[\frac{P(\boldsymbol{Y})}{P_e(\boldsymbol{Y})}\right] P_e\left(\boldsymbol{Y} \mid T_j = t\right) \cdot P_e\left(T_j = t\right),\end{aligned}$$

where $T_j \sim \mathcal{B}(n, t)$, then the above equation is continued:

$$= \binom{n}{t} p^t (1-p)^{n-t} \sum_{y_{ij}:T_j=t} \left[\frac{P(\boldsymbol{Y})}{P_e(\boldsymbol{Y})}\right] P_e\left(\boldsymbol{Y} \mid T_j = t\right),$$

where $P_e\left(y_{ij} \mid T_j = t\right) = \dfrac{1}{\binom{n}{t}}$, $y_{ij} \sim \text{Ber}(p_i)$ with $p_i = \dfrac{e^{\alpha_j \theta_i - \beta_j}}{1 + e^{\alpha_j \theta_i - \beta_j}}$ and p is the geometric mean of p_i.

Then we can use Monte Carlo method to calculate the weight; specifically, just sample items from uniform distribution M times and take average of those values as estimation of $\Pr(T_i = t)$.

References

[1] Daniel M. Bolt, Allan S. Cohen and James A. Wollack. A Mixture Item Response Model for Multiple-Choice Data. *Journal of Educational and Behavioral Statistics*, Vol. 26, No. 4 (2001), pp. 381-409.

[2] Jean-Paul Fox. *Bayesian Item Response Modeling: Theory and Application*. Springer, 2010.

[3] Kei Myazakli and Takahiro Hoshino. A Bayesian Semiparametric Item Response Model with Dirichlet Process Priors. *Psychometria*, Vol. 74, N0. 3 (2009), pp. 375-393.

[4] Leo A. Goodman. Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, Vol. 61, No.2 (1974), pp. 215-231.

[5] Maria Kalli, Jim E. Griffin and Stephen G. Walker. Slice sampling mixture models. *Statistics and Computing*, Vol. 21, No. 1 (2009), pp. 93-105.

[6] Michael D. Escobar and Mike West. Bayesian Density Estimation and Inference using Mixtures. *Journal of the American Statistical Association*, Vol. 90, No. 430 (1995), pp. 577-588.

[7] Noel Cressie and Timothy R.C. Read. Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 46, No.3 (1984), pp. 440-464.

[8] Stephen G. Walker. Sampling Dirichlet Mixture Model with Slices. *Communications in Statistics--Simulation and Computation*, Vol. 36, No.1 (2007), pp. 45-54.

[9] W. R. Gilks and P. Wild. Adaptive Rejection Sampling for Gibbs Sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 41, No. 2 (1992), pp. 337-348.