Three Contributions to Latent Variable Modeling

Xiang Liu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

ABSTRACT

Three Contributions to Latent Variable Modeling

Xiang Liu

The dissertation includes three papers that address some theoretical and technical issues of latent variable models. The first paper extends the uniformly most powerful test approach for testing person parameter in IRT to the two-parameter logistic models. In addition, an efficient branch-and-bound algorithm for computing the exact p-value is proposed. The second paper proposes a reparameterization of the log-linear CDM model. A Gibbs sampler is developed for posterior computation. The third paper proposes an ordered latent class model with infinite classes using a stochastic process prior. Furthermore, a nonparametric IRT application is also discussed.

Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I want to express my deep gratitude for Dr. Lawrence DeCarlo and Dr. Matthew Johnson, my advisers. I am forever indebted to your patient and careful guidance. I have lost count of how much time you spent advising me. Dr. Johnson, thank you for giving me the freedom and guiding me exploring many research topics. Our discussions are some of my most memorable moments during my graduate school career. Dr. DeCarlo, thank you for always sharing your perspectives. Your passion for psychometrics research encourages me to become a better researcher. I will always look up to my advisers not only as intellectuals but also as persons.

I want to thank other professors whom I had the honor to work with, especially Dr. Young-Sun Lee and Dr. Bryan Keller. Thank you for all the opportunities you brought me and research ideas you shared. I would also like to thank all the colleagues I worked with at EdLab. A special thanks goes to Dr. Gary Natriello and Dr. Hui Soo Chae. I am grateful for your support in all these years. I would not have made this far without your help. I would like to show my appreciations to Dr. Zhiliang Ying for being on my dissertation committee and sharing his insightful and valuable suggestions.

Lastly, I am thankful to my family. Thank you for the unconditional support you gave

me and the belief you had in me all these years. Especially, I want to thank my wife, Fei, for accompanying me through this journey.

To my grandfather

Chapter 1

Introduction

My research interests span the areas of latent variable models, categorical data analysis, and Bayesian methods. A common theme of my research is to understand and improve the theory and applications of psychometric models. To achieve this, I utilize techniques from both frequentist and Bayesian traditions. Specifically, a significant portion of my research deals with the development of computational methods for various psychometric models. Given the rapid improvement of computing power and the emergence of big data, it is certainly an area that will grow more relevant and more important. A wide range of popular psychometric models - item response theory (IRT), cognitive diagnosis models (CDM), latent class models have been covered in my research. Many of them have a Bayesian and computational focus.

1.1 The First Study: Small Sample Size Confidence Intervals for IRT

IRT is widely used in educational and psychological testing. One of the core purposes of IRT is to map students' ability onto a latent continuum. Since only a finite number of items can be administered, abilities are estimated with uncertainty. Traditionally, the standard error of the ability estimator is based on the large sample approximation (i.e., square root of the inverted fisher information). While it might be reasonable for longer tests, in reality, we often have to use short to medium length instruments (e.g., personality tests). The standard errors and the confidence intervals based on the large sample approximation could be highly inaccurate under these circumstances.

To address this important issue, the first study (X. Liu, Han, & Johnson, 2018) proposed a framework to construct hypothesis testing based on exact distribution. Confidence intervals of ability estimates are obtained by inverting the hypothesis tests . As a result, the type-I error rate is well controlled under even small to medium test lengths. A major hurdle to this approach is the heavy computational requirements. Instead of permuting all possible response patterns by brute force, I developed a branch and bound algorithm that can calculate p-values efficiently. With the help of the algorithm, the exact distribution approach is now computationally feasible for even medium test lengths. This work is not only technically interesting, but also enables practitioners and researchers to recognize the measurement uncertainty more accurately so that, ultimately, decision making can benefit from the improved measurement practice.

1.2 The Second Study: MCMC for Log-linear CDM Models

Bayesian statistics is another major piece of my research. Markov chain Monte Carlo (MCMC) has been widely used to estimate many kinds of psychometric models. Cognitive diagnosis models (CDM) is no exception. The second study (X. Liu & Johnson, n.d.) introduced a Gibbs sampler for estimating the saturated log-linear CDM model. By reparameterizing the log-linear CDM model, I was able to analytically derive the closed form update steps for the Gibbs sampler. The automatic update steps do not require tunning which makes it easy to use. I also gave the linear transformation that would transform the posterior samples back to the original log-linear CDM parameterization.

The introduced method potentially provides an automatic solution for researchers who may be interested in performing Bayesian analysis in CDM. Even though I introduced the method in the context of the saturated log-linear models, it can be easily extended to other specific CDM models.

1.3 The Third Study: Bayesian Nonparametric Ordered Latent Class Model

The third study develops Bayesian nonparametric methods for ordered latent class models. In latent class or mixture models, the number of classes usually has to be specified a priori. Essentially, it becomes a modeling choice that is either based on a researcher's substantive knowledge of the data or comparing the fit of models with different number of classes.

Selecting models based on some model fit indices is not always straightforward. Different fit indices might penalize the complexity of the model differently which may result in different conclusions. In addition, in some cases, a researcher might have to fit a large number of models with different number of classes before the optimal model can be determined. Therefore, it may create heavy computational burdens. More importantly, all inferences are conditioned on the selected model. It ignores the uncertainty of the model choice. By assigning stochastic process priors to latent class assignments, Bayesian nonparametrics can fit a model with an infinite number of classes. The posterior distribution of the number of classes provides a better picture of model uncertainty. Unlike the traditional methods, some inferences do not have to be conditioned on one selected model. Instead, marginal quantities can be obtained by averaging over the posterior distribution of the models with different dimensions which can account for the model uncertainty.

Chapter 2

The UMP Exact Test and the Confidence Interval for Person Parameters in IRT Models

2.1 Introduction

In Item Response Theory (IRT), the person parameter is often estimated with the maximum likelihood estimator (MLE) (Hambleton & Swaminathan, 1985). Under large sample sizes, the MLE is approximately normally distributed with asymptotic variance given by the inverse of the Fisher information (Baker & Kim, 2004). Based on the asymptotic normality, one can construct hypothesis tests and confidence intervals for the person parameter (Casella & Berger, 2001). However, the propriety of the asymptotic assumption is questionable under practical situations where the test is often of moderate lengths. As a result, the statistical inference of the person parameter based on the asymptotic normality of the MLE

could be very misleading. This problem has been recognized in earlier research (Lord, 1983; Klauer, 1991; Doebler, Doebler, & Holling, 2012; Biehler, Holling, & Doebler, 2014). Thus, developing statistical inference procedures that do not depend on the asymptotic normality is of practical importance. Two approaches are generally discussed.

One approach is to base the inference on the exact distribution of the response patterns. The idea of using the exact distribution of response patterns in IRT was initially introduced for the purpose of assessing response pattern fit (Molenaar & Hoijtink, 1990). Klauer (1991) further derived the uniformly most powerful unbiased (UMPU) test and the uniformly most accurate confidence interval based on the exact distribution of the response patterns in the Rasch model. Klauer noticed that the raw sum score is a sufficient statistic for the person parameter in the Rasch model. Therefore, response patterns in the Rasch model can be reduced to raw sum scores. Furthermore, the test statistic based on the raw sum score was randomized to achieve the level α *unbiased* test. The computation is tractable as only the exact distribution of the raw sum scores are needed. As the number of items increases, the number of possible raw sum scores increases linearly.

The exact distribution approach has also been extended to the two-parameter logistic (2PL) model. Unfortunately, the raw sum score is no longer a sufficient statistic for the person parameter in the 2PL model. Instead of calculating the exact distribution of the sufficient statistic, the exact distribution for a Wald statistic was calculated (Doebler et al., 2012). And confidence intervals were derived by inverting the exact test. In the same paper, the authors also proposed a hybrid Bayesian approach by incorporating a prior distribution on the person parameter.

A second class of approaches is based on approximating the exact distribution of the

person parameter using a higher-order approximation (Biehler et al., 2014). The saddle-point approximation works well for IRT models within the exponential family and is relatively easy to implement. However, it does not yield the optimal confidence interval such as the uniformly most accurate confidence interval (Klauer, 1991).

Given the much improved computing power, calculating the exact distribution under small to moderate item lengths becomes feasible. In the present work, we extend Klauer's (1991) approach to the 2PL model. In fact, we generalize the procedure for IRT models in the exponential family. In addition, an efficient branch and bound algorithm is introduced.

2.2 Existing Methods

The small sample inference methods mentioned in the introduction have been rarely discussed in the IRT literature. As a result, readers may not be familiar with them. We will briefly review these ideas in this section.

2.2.1 Saddle-point Approximation

A probability distribution can be completely characterized by its characteristic function (Casella & Berger, 2001). The density function of a distribution can be obtained by inverse-Fourier transformation of the characteristic function. Often, the transformation has to be approximated. The goal of the saddle-point approach here is to provide a robust approximation to the distribution of the sufficient statistic given an ability parameter, i.e. $P\left(T\left(X\right)|\theta_0\right)$, under small sample size. Biehler et al. (2014) described the approximation in the appendix of their paper. The derivation involves exponential tilting and Edgeworth expansion. Com-

pared to an error term of $O(n^{-1/2})$ from the first-order normal approximation, the saddle-point approximation has an error term of $O(n^{-1})$. Consequently, it converges faster to the true distribution as the number of items increases (Biehler et al., 2014). This desired feature provides a relatively accurate approximation under small sample sizes. The tail probabilities can be obtained through integrating the approximated probability density function. The approximation of this integral is provided by the Lugannani-Rice formula (Lugannani & Rice, 1980). Then inverting two equal tail tests gives the confidence interval.

The approximation depends on the availability of the mean and the variance of the distribution of the sufficient statistic. This is possible for the exponential family as the distribution of the sufficient statistic is also within the same family. The cumulant generating function of the sufficient statistic is known, and its mean and variance are given by the first two cumulants.

This method only works for IRT models within exponential family. Although the approximation is highly accurate, simulations show that the coverage rate of the resulting confidence interval may still fall below the nominal rate for some conditions (Biehler et al., 2014).

2.2.2 Exact Distribution Approaches

The methods proposed in Doebler et al. (2012) are more closely related to our approach in the sense that they are all based on the exact finite sample distribution of the response patterns.

2.2.2.1 The Hybrid Bayesian Approach

To test the two-sided hypothesis $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$, Doebler et al. (2012) proposed a likelihood ratio type statistic in the form of

$$l_{\theta_0}(\boldsymbol{x}) = \frac{P_f(\boldsymbol{x})}{P_{\theta_0}(\boldsymbol{x})}, \tag{2.1}$$

where $P_f(\boldsymbol{x}) = \int_{\theta \neq \theta_0} P_\theta(\boldsymbol{x}) f(\theta) d\theta$, $P_\theta(\boldsymbol{x})$ is the probability of the response vector $\boldsymbol{x}$ under some ability level θ, and $f(\theta)$ is the prior distribution of the person parameter. Intuitively, l in equation 2.1 is a ratio of the weighted average of likelihood over a prior distribution to the likelihood under the hypothesized value $\theta = \theta_0$. A larger l would provide stronger evidence to reject θ_0. Then the acceptance region can be defined as

$$A_{1,\theta_0} = \{\boldsymbol{x} : l_{\theta_0}(\boldsymbol{x}) \leq C\}.$$

The constant C is chosen such that it is the smallest that can satisfy $P_{\theta_0}(A_{\theta_0}) \geq 1 - \alpha$ for some nominal level α. Due to discreteness, the nominal level generally cannot be achieved exactly.

The associated confidence set can be obtained by inverting the above test. The idea is to find all θ that will not be rejected given some observed response pattern. So the confidence set is

$$I_1(\boldsymbol{x}) = \{\theta : \boldsymbol{x} \in A_{1,\theta}\}. \tag{2.2}$$

Doebler et al. (2012) showed that the interval in (2.2) minimizes the average expected length,

$$\int_\theta E_\theta \mu(I)) f(\theta) d\theta,$$

among confidence sets with the same level of significance.

2.2.2.2 The Exact Normal Approach

The idea of this approach is to construct confidence intervals based on the asymptotic variance of the estimator. Instead of getting probabilities from the asymptotic distribution, the authors calculated probabilities exactly from the finite sample distribution (Doebler et al., 2012). For the same two-sided hypothesis, a Wald-type statistic was proposed,

$$z_{\theta_0}(\boldsymbol{x}) = \frac{\hat{\theta}(\boldsymbol{x}) - \theta_0}{\sqrt{var\left(\hat{\theta}(\boldsymbol{x})\right)}}, \tag{2.3}$$

where $\hat{\theta}$ is an estimator for θ. Note that $\hat{\theta}$ does not have to be the MLE. Other estimators such as weighted likelihood estimator (Warm, 1989) would also work here. Similar to the Hybrid Bayesian approach, the acceptance region is then

$$A_{2,\theta_0} = \{|z_{\theta_0}| \leq C\}.$$

The constant C is the smallest that can satisfy $P_{\theta_0}(A_{\theta_0}) \geq 1 - \alpha$ for some nominal level α. The confidence set can be obtained by inverting the test in a similar fashion.

$$I_2(\boldsymbol{x}) = \{\theta : \boldsymbol{x} \in A_{2,\theta}\}.$$

This approach is similar to the standard approach based on asymptotic normality. However, the authors did not state any optimality for this method (Doebler et al., 2012).

2.2.2.3 Limitations

This class of approaches does not rely on approximation of the distribution of a test statistic. Instead, probabilities are calculated based on the exact distribution of response patterns. As a result, the coverage rate of the confidence set will never fall below the nominal level.

But there are limitations. Due to discreteness, the coverage rate of the confidence set might be higher than the nominal level. In other words, confidence sets from methods based on exact distribution are often conservative. In addition, the Bayesian approach is sensitive to the choice of prior. The exact normal approach does not generally meet any optimality criterion. More importantly, the resulting confidence *sets* from the above two methods are not necessarily *intervals* (Doebler et al., 2012). It is due to the fact that the endpoints of the acceptance region need not be monotone in θ_0 (Agresti, 2003). Casella and Berger (2001) also discussed this problem.

The exact distribution approach using sufficient statistic (Klauer, 1991) avoids this particular problem. Furthermore, it leads to the uniformly most powerful test which is a more common optimality criterion.

2.3 Theory

In this section, we derive the Uniformly Most Powerful (UMP) test for IRT models in the exponential family.

2.3.1 IRT models in the exponential family

Lord (1980) showed that 1PL and 2PL models with known item parameters are in the exponential family. Here, without loss of generality, we demonstrate the result for the 2PL model when the item parameters are known. The item response function for the 2PL model is given by

$$P_j(X_j = 1|\theta_i) = \frac{\exp(a_j(\theta_i - b_j))}{1 - \exp(a_j(\theta_i - b_j))},$$

where θ_i is the latent ability of i^{th} subject, a_j is the discrimination parameter for j^{th} item, and b_j is the difficulty parameter for j^{th} item. Then for a given response pattern $\boldsymbol{x}$, the likelihood function for θ can be written as

$$L(\theta|\boldsymbol{x}) = \prod_{j=1}^{n} P_j^{x_j}(1-P_j)^{1-x_j} = \prod_{j=1}^{n} \left\{ \frac{\exp[a_j(\theta - b_j)]}{1-\exp[a_j(\theta-b_j)]} \right\}^{x_j} \left\{ \frac{1}{1-\exp[a_j(\theta-b_j)]} \right\}^{1-x_j}.$$

The likelihood in equation 2.3.1 can be further factorized into the following form (for details, see appendix)

$$L(\theta|\boldsymbol{x}) = \exp[\eta(\theta)T(\boldsymbol{x})]h(\boldsymbol{x})g(\theta),$$

where

$$\exp[\eta(\theta)T(\boldsymbol{x})] = \exp(\theta \sum_{j=1}^{n} a_j x_j),$$
$$h(\boldsymbol{x}) = \exp(-\sum_{j=1}^{n} a_j x_j b_j),$$

and

$$g(\theta) = \prod_{j=1}^{n} \{1-\exp[a_j(\theta-b_j)]\}^{-1}.$$

This shows that 2PL model is in the exponential family and θ is a natural parameter. Furthermore, $T(\boldsymbol{x}) = \sum_{j=1}^{n} a_j x_j$ is a *sufficient statistic* for θ. Under the 1PL model, the sufficient statistic reduces to the raw sum score, since the discrimination parameters are constant across items. Furthermore, because $\eta(\theta) = \theta$ is an increasing function, 1PL and 2PL models have the monotone likelihood ratio (MLR) property (Casella & Berger, 2001).

2.3.2 The UMP one-sided hypothesis test

Consider testing $H_0 : \theta \geq \theta_0$ *versus* $H_1 : \theta < \theta_0$ with a test of the form

$$\phi_1 = \begin{cases} 1 & T(\boldsymbol{x}) < C \\ \\ 0 & T(\boldsymbol{x}) \geq C, \end{cases}$$

where $T(\boldsymbol{x}) = \sum_{j=1}^{n} a_j x_j$ is a sufficient statistic for θ. If $\phi = 1$, the test rejects H_0. If $\phi = 0$, the test fails to reject H_0. The constant C is minimal that can satisfy $P_{\theta_0}(T(\boldsymbol{X}) < C) \leq \alpha$, for some nominal level α. Because $T(\boldsymbol{x})$ is a sufficient statistic for θ and the MLR property holds, by the *Karlin-Rubin* theorem, the test ϕ is a UMP level $\alpha^\star = P_{\theta_0}(T(\boldsymbol{X}) < C)$ test (Casella & Berger, 2001).

In the context of 1PL and 2PL models, the distribution of the sufficient statistic is discrete. Thus, in general, $\alpha^\star$ cannot achieve the nominal level α. If a size α test is desired, the sufficient statistic can be randomized by adding a random component (Klauer, 1991). However, the randomized statistic leads to randomized decision. In practice, randomized decisions are problematic. For example, two examinees who have the exact same response pattern might receive different decisions. This raises a serious fairness question. In this paper, we do not provide details for randomizing the test.

Instead of calculating the constant C, the exact p-value can also be obtained. For a given observed response pattern $\boldsymbol{x}_0$, the p-value is given by

$$P_{\theta_0}(T(\boldsymbol{X}) \leq T(\boldsymbol{x}_0)) = \sum_{\boldsymbol{x}:\ T(\boldsymbol{x}) \leq T(\boldsymbol{x}_0)} P_{\theta_0}(\boldsymbol{X} = \boldsymbol{x}). \tag{2.4}$$

For testing $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, we can compare the exact p-value against the nominal level. If the p-value is less than the nominal level, we reject H_0; otherwise, we fail to reject H_0.

In this section, we derived the one-sided UMP test for IRT models in the exponential family. Notice that for testing $H_0 : \theta \leq \theta_0$ versus $\theta > \theta_0$, the derivation of the UMP test follow the same approach.

2.3.3 The two-sided test

For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, the UMP test does not exist (Casella & Berger, 2001; Klauer, 1991). Klauer (1991) was able to derive the two-sided UMP unbiased test. However, this approach requires a continuous test statistic from randomization. Here, we construct an equal-tail two-sided test with some nominal level α. The test has the following form

$$\phi_2 = \begin{cases} 1 & T(\boldsymbol{x}) < C_1 \text{ or } T(\boldsymbol{x}) > C_2 \\ \\ 0 & C_2 \leq T(\boldsymbol{x}) \leq C_1, \end{cases}$$

where $T(\boldsymbol{x}) = \sum_{j=1}^{n} a_j x_j$. C_1 is maximal such that

$$P_{\theta_0}\left(T\left(\boldsymbol{x}\right) < C_1\right) \leq \alpha/2. \tag{2.5}$$

Similarly, C_2 is minimal that satisfies

$$P_{\theta_0}\left(T\left(\boldsymbol{x}\right) > C_2\right) \leq \alpha/2. \tag{2.6}$$

From this definition, one can see that the test is essentially two one-tailed tests with equal significance level $\alpha/2$. Again, due to discreteness, in general the equal signs in (2.5) and (2.6) cannot be achieved.

2.3.4 The confidence interval

Inverting the equal tail test ϕ_2 leads to the confidence set. Given an observed response pattern $\boldsymbol{x}_0$, we find all θ that will not be rejected,

$$I_3(\boldsymbol{x}_0) = \{\theta : \phi_{2,\boldsymbol{x}_0}(\theta) = 0\}.$$

Unlike I_1 or I_2, the endpoints of the acceptance region in this method is monotone in θ_0. This guarantees that the set I_3 is an *interval*. This is also referred to as the tail method (Agresti, 2003). Consequently, the lower bound of the confidence interval can be obtained by solving

$$P_{\theta_0}(T(\boldsymbol{X}) \leq T(\boldsymbol{x}_0)) = \alpha/2,$$

for θ_0. Similarly, the solution for

$$P_{\theta_0}(T(\boldsymbol{X}) \geq T(\boldsymbol{x}_0)) = \alpha/2, \tag{2.7}$$

gives the upper bound. Now I_3 can be expressed as

$$I_3(\boldsymbol{x}_0) = \{\theta : \theta_l \leq \theta \leq \theta_u\}, \tag{2.8}$$

where θ_l and θ_u are solutions of (2.7) and (2.8) as functions of θ, which can be obtained by standard root finding algorithms.

2.3.5 Computational algorithm

One difficulty associated with the exact distribution approach is the computational complexity. For the 1PL model, the computation is manageable. As the number of items increases, the number of possible raw sum scores increases linearly. However, under the 2PL model,

the number of weighted sum scores, in general, increases exponentially. Therefore, computation by brute force would quickly become too time consuming to be feasible as the test length increases. In this section, we introduce an efficient branch and bound algorithm for calculating exact p-values.

2.3.5.1 Formulation of the problem

By brute force, finding the exact p-value of a given response pattern $\boldsymbol{x}_0$ as defined in (2.4) requires the complete enumeration of all response patterns. However, it should be noticed that only response patterns that produce sufficient statistics not larger than the sufficient statistic of the observed response pattern need to be considered. Thus, the problem translates to finding all possible response pattern $\boldsymbol{x}$ such that

$$a_1x_1 + a_2x_2 + \cdots + a_jx_j \leq T(\boldsymbol{x}_0). \tag{2.9}$$

Enumerating subject to a constraint in the form of (2.9) is a discrete mathematical programming problem that can be solved by the branch and bound algorithm. The algorithm was first conceptualized in 1960 (Land & Doig, 1960), and was later formalized and given its current name for the purpose of solving the well-known traveling salesman problem (Little, Murty, Sweeney, & Karel, 1963).

2.3.5.2 The binary tree representation

The response patterns can be represented using a binary tree structure. Figure 21 shows an example of three items. Each branch of the tree represents a response pattern. For example, the branch $0 \rightarrow 0 \rightarrow 0$ represents the response pattern $(0, 0, 0)$. For J items, there are 2^J

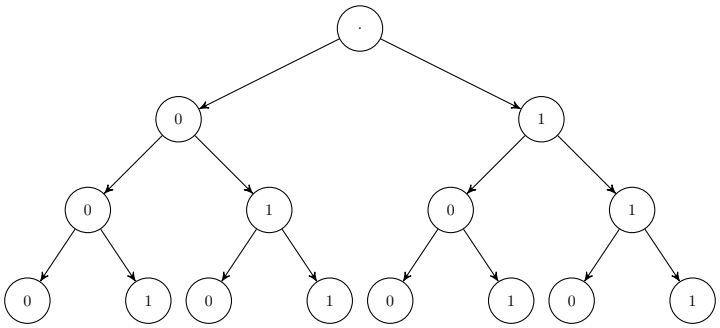


Figure 21: The binary tree representation of response patterns

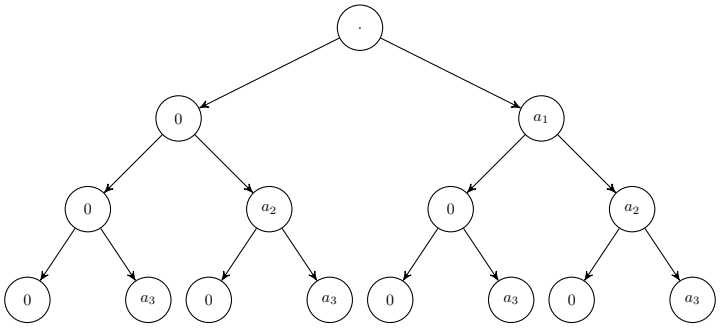


Figure 22: The binary tree representation of weighted sum scores

possible response patterns. Thus, the number of the nodes at the bottom of the tree is 2^J. The tree has $J + 1$ levels, with $(j + 1)^{th}$ level representing the response for j^{th} item.

Each response pattern is associated with a weighted sum score. Therefore, the weighted sum scores can also be represented using the same binary tree structure by multiplying together the dichotomous responses in Figure 21 and the associated discrimination parameters (see Figure 22). Adding the nodes within a branch leads to the weighted sum score for the response pattern presented by the branch. For example, the weighted sum score for the response pattern $(1, 0, 1)$ is $a_1 + 0 + a_3$.

2.3.5.3 Branch and bound

Because the item discrimination parameters in the 2PL are usually constrained to be positive, the weighted sum score is non-decreasing as we go deeper along any branch in the tree. For a given observed response pattern $\boldsymbol{x}_0$, the constraint $T(\boldsymbol{x}_0)$ in (2.9) can be computed. Then we only need to consider those branches that satisfy the constraint.

The construction of the binary tree can be viewed as a series of splitting operations. While the constraint is satisfied, the branch keeps splitting. The weighted sum score is checked against the constraint after each splitting. If the constraint is not satisfied, any further splitting could not possibly produce response patterns that could satisfy (2.9) due to the non-decreasingness mentioned above. Thus, the branch can be terminated.

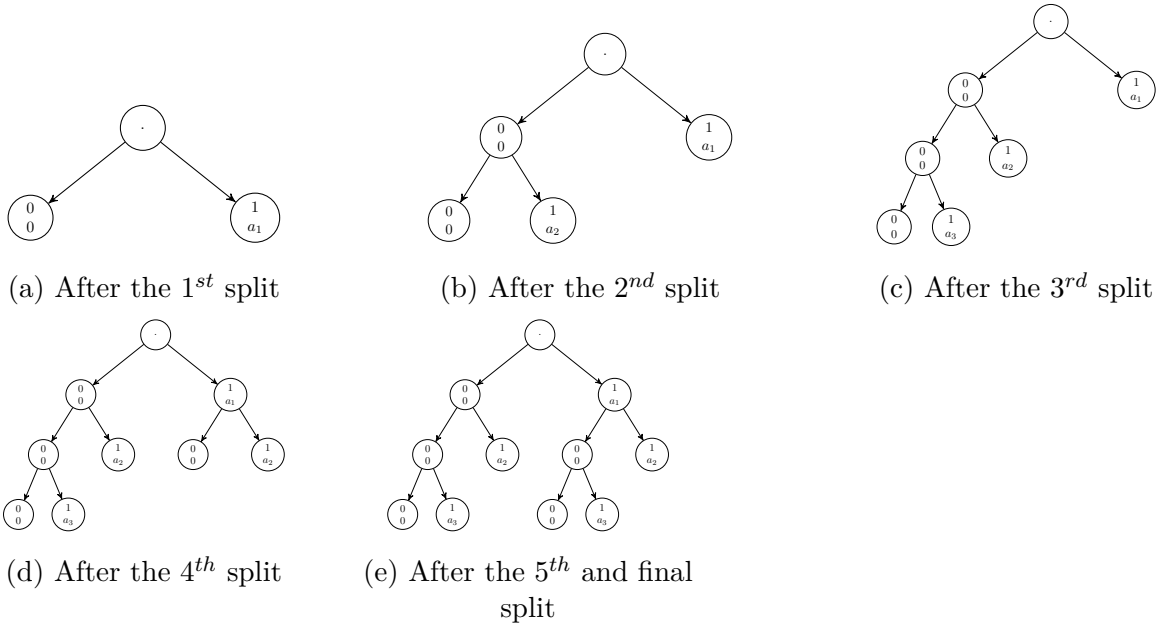
For example, suppose $\boldsymbol{x}_0 = (1, 0, 0)$ and $a_1 < a_2 < a_3$. The sufficient statistic of the observed response pattern is then $T(\boldsymbol{x}_0) = a_1$. The splitting process is demonstrated in Figure 23.

Notice that, after the 3^{rd} split, the branch $0 \to 1$ did not split any further as the weighted sum is $0 + a_2 > a_1$. After the final split, all the response patterns that could potentially satisfy the constraint (2.9) should have reached the bottom of the tree. Out of the four branches reached the bottom, only $0 \to 0 \to 0$ and $1 \to 0 \to 0$ satisfy (2.9). Thus the exact p-value is

$$P_{\theta_0}(T(\boldsymbol{X}) \leq T(\boldsymbol{x}_0)) = P_{\theta_0}\{\boldsymbol{X} = (000)\} + P_{\theta_0}\{\boldsymbol{X} = (100)\}. \tag{2.10}$$

Instead of calculating probabilities for all eight possible patterns, only two probabilities need to be evaluated using the branch and bound algorithm.

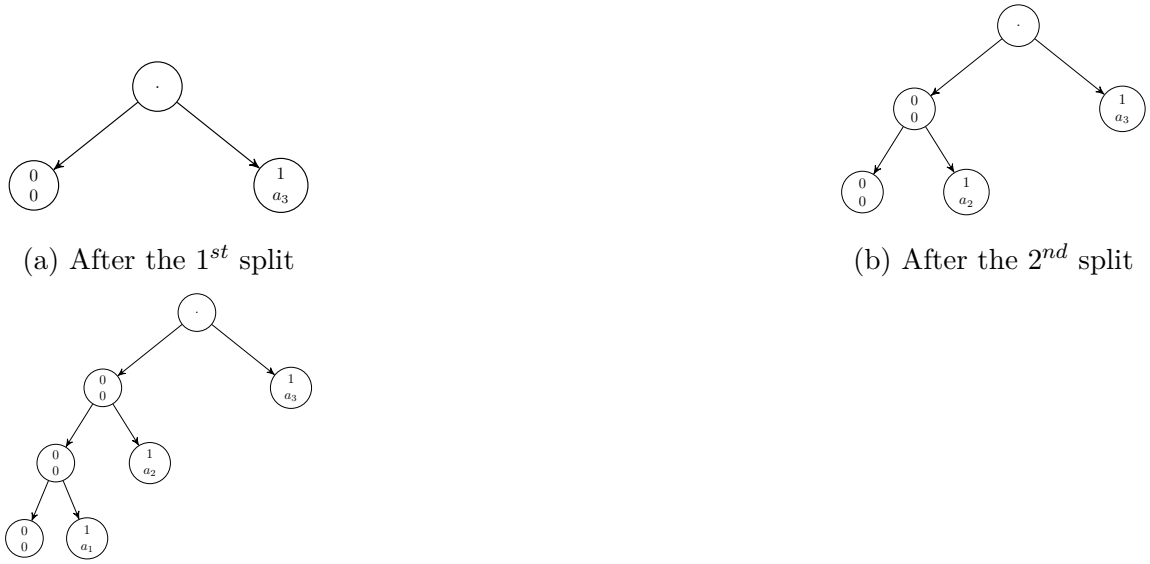
The efficiency of the algorithm can be further improved by ranking the items in descend-



(a) After the 1^{st} split (b) After the 2^{nd} split (c) After the 3^{rd} split

(d) After the 4^{th} split (e) After the 5^{th} and final split

Figure 23: Demonstration of the splitting process



(a) After the 1^{st} split (b) After the 2^{nd} split

(c) After the 3^{rd} split

Figure 24: The splitting process for ranked items

ing order according to their discrimination parameters before constructing the tree. For the same example above, the splitting process after ranking is demonstrated in Figure 24. Compared to the five splits in Figure 23, it only needs three splits after ranking the items.

However, we should take the computational cost of ranking into our consideration. But in our experience, the cost of ranking is negligible compared to the improved efficiency in the branch and bound algorithm.

In the example we provided, the tree is always traversed as deep as possible along a branch before exploring parallel branches. This is called the *Depth First Search*(DFS) (Leiserson, Rivest, Stein, & Cormen, 2009). We implemented this version of the branch and bound algorithm for calculating exact p-values. We also used Brent's algorithm (Brent, 1973) to find bounds for the confidence interval. The C++ code is available from the first author.

2.4 Simulation Study

2.4.1 Type-I error and power of the one-tail test

A simulation study was conducted to examine the type-I error rate and the power of the proposed exact test for the 2PL model under different conditions. The difficulty with such a simulation study is that the power of the test is associated with many different factors. The number of items is expected to affect the power. Also, for a given set of items, the power is likely to vary for different ability levels. Moreover, the power depends on the difference between the *tested* ability level and the *true* ability level, i.e.

$$\Delta\theta = \theta_0 - \theta. \tag{2.11}$$

However, the most difficult part is perhaps that different sets of item parameters would likely to have different power curves. Considering this, we decided to investigate the average power for item parameters from their usual distributions under different conditions.

2.4.1.1 Design

Four test lengths were considered: $J = 5, 10, 15$, and 20 items. For each test length, seven effect sizes as defined in Equation 2.11 were examined, $\Delta\theta = 0, 0.5, 1.0, \ldots, 3.0$. Moreover, twenty-five hypothesized ability level θ_0s were selected from -3 to 3 by intervals of 0.25. Then the true ability level θs were computed by $\theta = \theta_0 - \Delta\theta$.

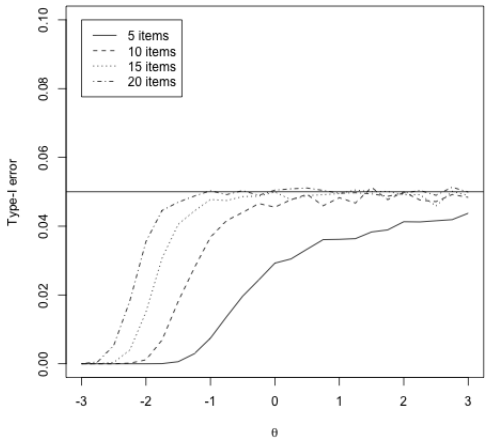
Given a pair of J and $\Delta\theta$, item difficulty parameters b_j and item discrimination parameters a_j, were randomly generated from uniform distributions, $U(-2.0, 2.0)$ and $U(0.5, 2.0)$ respectively. At each θ, a response pattern $\boldsymbol{x}_0$ was randomly generated using item parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ under the 2PL model. Then we test the one-sided hypothesis $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$. The exact p-value was calculated as in (2.4) using the branch and bound algorithm. H_0 is rejected if the p-value is less than $\alpha = 0.05$.

For each combination of the length J and the effect size $\Delta\theta$, the above process was replicated $50,000$ times.

2.4.1.2 Results

Figure 25 shows the average type-I error rate ($\Delta\theta = 0$) under different item lengths. When there are only $J = 5$ items, the type-I error rate is significantly lower than the nominal level $\alpha = 0.05$. This is due to the discreteness of the response patterns. For 5 items, there are $2^5 = 32$ possible response patterns. However, as the number of items increases, the number of possible response patterns increases exponentially. As a result, the type-I error rate is getting very close to the nominal level for moderate and high θ_0 when item length increases even without randomization. The power drops eventually to zero as the θ_0 moves to the

Figure 25: The type-I error rate from the simulation

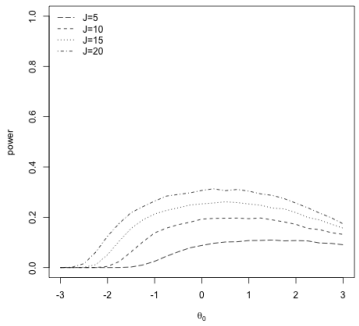


extreme negative value. For a given set of items, the response pattern $\boldsymbol{x} = (0, 0, \ldots, 0)$ has the weighted sum score $\sum_{j=1}^{J} a_j x_j = 0$ which is the smallest among all possible response patterns. If θ_0 is so small that $P_{\theta_0}(\boldsymbol{X} = (0, 0, \ldots, 0)) > \alpha$, the test will not reject θ_0 given any observed response pattern at the nominal level α. As we pointed out earlier, if the nominal level is desired, the decision has to be randomized.

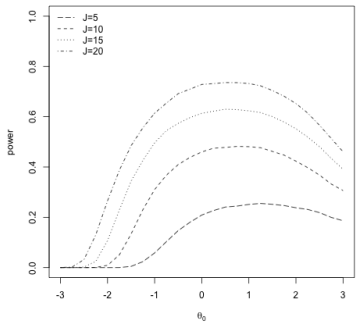
The power curves for $\Delta\theta = 0.5, 1.0, \ldots, 3.0$, are shown in Figure 26. Power is highest for null values of ability, θ_0, in the middle part of the scale. Power drops as the null value, θ_0, moves towards the two extremes. The shape of the power curve is likely due to the way we generated the item parameters, which produced a test information curve that is highest near zero on the ability scale.

As expected, the power increases as the number of items increases. The power also increases with $\Delta\theta$. For a five item instrument, the power is quite low across all ability levels except for very large $\Delta\theta$. Meanwhile, for the item length $J = 20$, the power is reasonably good around medium level θ_0. When $\Delta\theta = 0.5$, the power is low, even for a twenty item

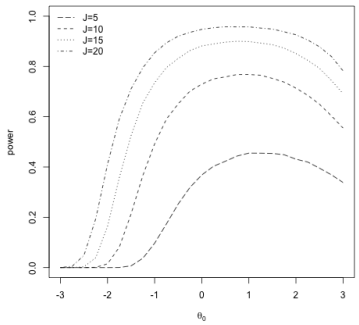
Figure 26: Statistical power under different conditions



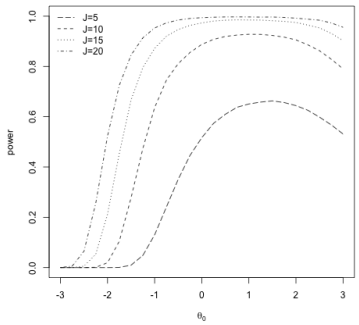
(a) $\Delta\theta = 0.5$



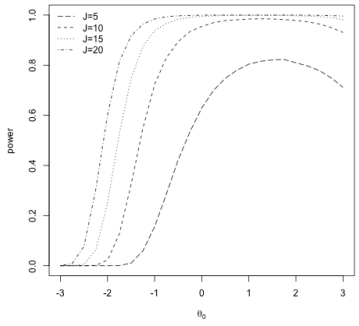
(b) $\Delta\theta = 1.0$



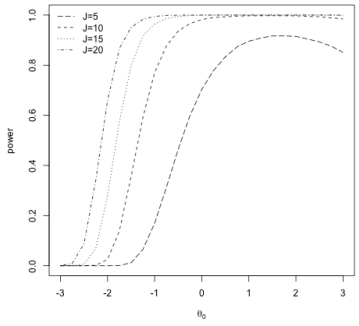
(c) $\Delta\theta = 1.5$



(d) $\Delta\theta = 2.0$



(e) $\Delta\theta = 2.5$



(f) $\Delta\theta = 3.0$

instrument.

2.4.2 Coverage rate of the confidence interval

2.4.2.1 Design

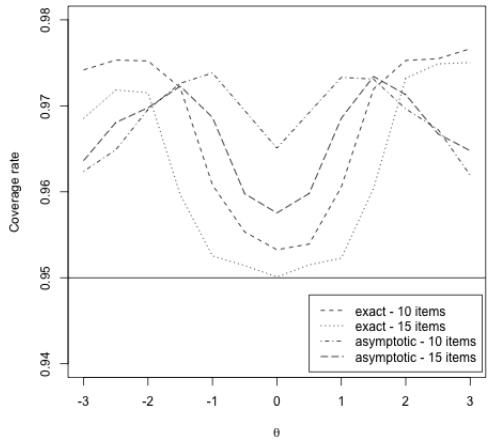
To examine the coverage rate of the proposed confidence interval, we conducted the second simulation study. Item discrimination parameters and difficulty parameters were generated from the same uniform distributions in the first simulation. We considered 13 ability parameters from -3 to $+3$ by intervals of 0.5. Response patterns were generated from the 2PL model. For each generated response pattern, the bounds of the 95% confidence interval were obtained by solving (2.7) and (2.8). In the case of extreme patterns, the bounds were limited to -5 and $+5$. Two item lengths were considered: 10 and 15. The process replicated $50,000$ times for each condition.

We also calculated confidence intervals based on the standard asymptotic approach and examined the coverage rates for comparison. For the 2PL model, the Fisher information is given by

$$\omega(\hat{\theta}) = \sum_{j=1}^{J} \frac{a_j^2 \exp(-a_j(\hat{\theta} - b_j))}{[1 + \exp(-a_j(\hat{\theta} - b_j))]^2}, \tag{2.12}$$

where $\hat{\theta}$ is the MLE of the person parameter. The asymptotic standard error then can be obtained by taking square root of the inverted Fisher information, i.e. $S.E. = \frac{1}{\sqrt{\omega(\hat{\theta})}}$ (Baker & Kim, 2004).

Figure 27: Coverage rate of the 95% confidence interval



2.4.2.2 Results

Figure 27 shows the coverage rate of the 95% confidence interval for different test lengths. As expected from the exact approach, the coverage rate does not fall below the nominal level for the entire range of ability parameter θ. When θ is close to 0, the coverage rate is very close to the nominal level. As the it goes towards extreme, the coverage rate is getting higher. Agresti (2003) noted this problem in the context of binomial proportions. When θ is sufficiently low, the interval can never exclude θ by falling below it. In this case, the lower bound on the coverage rate is actually $1-\alpha/2$ rather than $1-\alpha$. The other direction follows the same rationale. that being said, the coverage rate never exceeds 98% in our simulations even for extreme person parameters. More items would lead to a "less discrete" test statistic under the 2PL model. Consequently, the coverage rate is closer to the nominal level. For a test with 15 items, the coverage rate is very close to 95% for a good range of θ, and almost exact for medium level θ. On the other hand, the confidence intervals based on the standard asymptotic approach are overly conservative. Even with 15 items, the coverage rate is still

not close to the nominal level for medium level θ.

2.4.3 Lengths of the confidence interval

2.4.3.1 Design

The purpose of this simulation study is to examine the lengths of the proposed confidence interval for different levels of generating person parameter. We also compared the results with the standard confidence interval based on the asymptotic normality of the MLE. We considered 5 ability parameters, i.e. $\theta = -3.0, -2.0, \ldots, 3.0$. For each θ, we generated 10000 responses from the 2PL model. Item parameters were treated as random effects and generated in the same fashion as in the previous simulations. Average length of the confidence intervals are computed for both the exact approach and the standard asymptotic approach.

2.4.3.2 Results

Table 21: Average confidence interval lengths

(a) 10 items

θ	exact	asymptotic
-3.0	3.83	9.62
-1.50	3.16	4.17
0.0	2.60	2.58
1.50	3.30	4.17
3.0	4.58	9.55

(b) 15 items

θ	exact	asymptotic
-3.0	3.37	6.62
-1.50	2.55	2.77
0.0	2.05	2.03
1.50	2.58	2.78
3.0	3.98	6.66

Table 21 shows average confidence interval lengths for both the exact distribution approach and the standard asymptotic approach under 10 items and 15 items. While maintaining adequate coverage rates as demonstrated in the previous simulation, the exact dis-

tribution approach results in shorter confidence intervals across different ability levels. The difference is smaller in the middle of the ability level. The average lengths of the proposed confidence interval are comparable to those reported in Doebler et al. (2012) in the medium θ level, but shorter for more extreme θ values.

2.4.4 Computational time

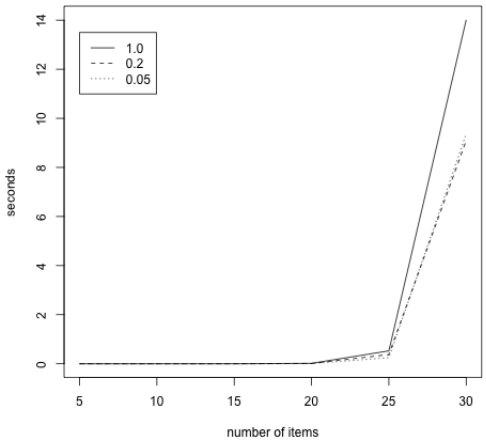
2.4.4.1 Design

In order to investigate the feasibility of our proposed method, we benchmarked computational time required. Six test lengths were considered - 5, 10, 15, 20, 25, and 30. Item parameters were generated from the same uniform distributions used in the other two simulations. Person parameters were generated from a standard normal distribution. Response patterns were generated under 2PL model. For the problem of testing the one-sided hypothesis $H_0 : \theta \geq 1.2$ versus $H_1 : \theta < 1.2$, the time required to compute an exact p-value was recorded. We ran 100 replications, and the average time was taken. We also benchmarked efficiency of finding confidence intervals for various test lengths.

2.4.4.2 Results

The average number of seconds required for computing exact p-values are in Figure 28. The p-values are calculated up to the specified thresholds - 1.0, 0.2, and 0.05. When the test length is not greater than 25, it takes well less than 1 second to compute an exact p-value. The time difference between different threshold is minimal. When the test length is 30, the computational cost becomes significantly higher. Now, computing an exact p-value takes

Figure 28: Average computing time for an exact p-value



about 14 seconds. If we are content with a threshold of 0.2 or 0.05, the average time needed would be roughly 9 seconds.

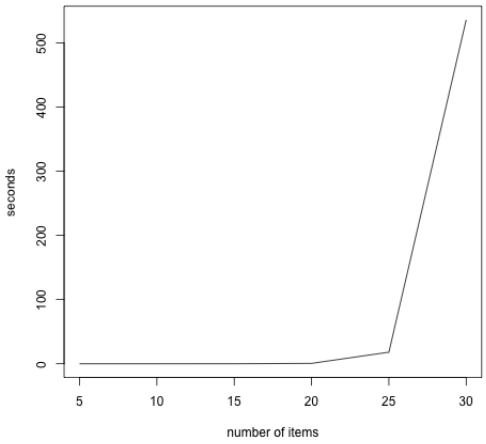
The average time required for computing a confidence interval under different test lengths are in Figure 29. Computing the confidence interval for a response pattern with no more than 20 items is very fast and takes less than 1 second. But for a 25-item pattern, it would take 18 seconds on average. When the test length is 30, the average computing time quickly increases to about 540 seconds.

2.5 Real data example

2.5.1 Hypothesis testing for LSAT data

In this section, we demonstrate a practical application of the proposed method. Dichotomous responses to 5 questions from 1000 individuals were extracted from the *R* package *ltm* (Rizopoulos, 2006). The data set is from the Law School Admission Test(LSAT) (Bock &

Figure 29: Average computing time for a confidence interval

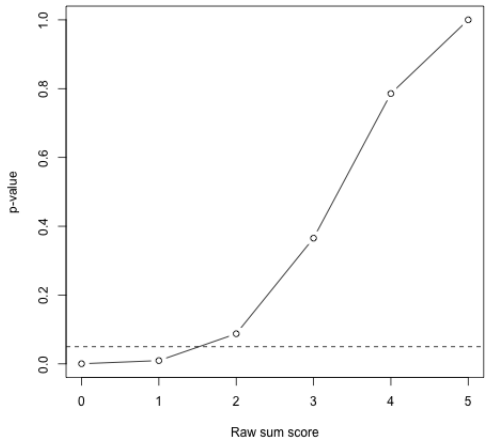


Lieberman, 1970).

In this example, we fit the Rasch model to the data. The difficulty parameters, $\beta_j, j = 1, 2, \ldots, 5$, are estimated through conditional maximum likelihood estimation using the R package *eRm* (Mair & Hatzinger, 2007). For identification purposes, we constrained the parameters so that $\sum_{i=1}^{5} \beta_j = 0$. The estimated parameters are then treated as known.

We are interested in testing the one-sided hypothesis $H_0 : \theta \geq 1.28$ against $H_1 : \theta < 1.28$. Under the Rasch model, the sufficient statistic for the ability parameter is the raw sum score. Therefore, $2^5 = 32$ possible response patterns can be reduced to 6 raw sum scores. The exact p-value for each raw sum score is calculated and presented in Figure 210. If the nominal level is set at $\alpha = 0.05$, H_0 will be rejected when the raw sum score is 0 or 1. For raw sum scores $2, 3, \ldots, 5$, we don't have enough evidence to reject H_0.

Figure 210: The exact p-values for the LSAT data

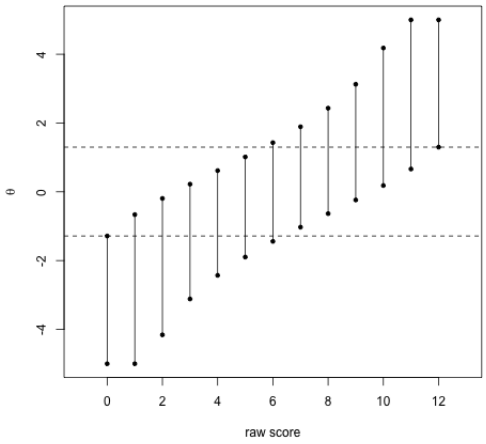


2.5.2 Confidence interval for the SF-12 data

The 12-item Short-Form Health Survey (SF-12) is perhaps one of the most widely used patient-reported heath outcome rating scales (J. E. J. Ware & Sherbourne, 1992; J. E. Ware, Kosinski, & Keller, 1996). Hagell and Westergren (2011) analyzed a data set of 150 Parkinson's disease patients responding to the SF-12 survey. The original SF-12 survey has 4 dichotomous items and 8 polytomous items. The authors (Hagell & Westergren, 2011) examined thresholds between adjacent response categories for the ploytomous items and decided there were too many categories. They proceeded to dichotomize item responses by collapsing adjacent categories and fitted the Rasch model. A reasonably good fit was reported.

One of the concerns of using short test forms is whether there is enough information to reliably distinguish subjects. Based on the parameter estimates obtained for the Rasch model (see Table 4 in Hagell & Westergren, 2011), we computed confidence intervals for each raw sum score (see Figure 211). In order to be significantly different from the lowest

Figure 211: 95% confidence intervals for the SF-12 data



score, a person has to obtain a raw score that is at least 7. Similarly, only raw scores less than 6 can be distinguished from the highest score at the nominal level. If a person scored 6 in the survey, there would not be enough evidence to distinguish this individual from any other person.

2.5.3 Hypothesis testing for the food security data

The 10 item food security data were extracted from the 2002 Current Population Survey and analyzed using IRT models (Johnson, 2004). The responses of 9804 individuals were used approximate the posterior distribution of the model parameters for both the Rasch and the 2PL models. The discrimination parameters varied significant. As a result, the Rasch model did not provide an adequate fit and 2PL model was favored. A cutoff score (1.93) was determined so that 95% of food insecure population with latent score greater than 1.93 will respond affirmatively to at least six of the food security items under the IRT model. For details of finding the cutoff score, please refer to Johnson (2004).

Based on the 2PL item parameters reported (see Table 1 in Johnson, 2004), we calculated exact p-values for the one-sided hypothesis $H_0 : \theta \leq 1.93$ against $H_1 : \theta \geq 1.93$. Of the possible $2^{10} = 1024$ patterns, 18 are rejected at the $\alpha = 0.05$ level (see Table 22 in Appendix B). Among them, 1 pattern has a raw score of 10, 10 patterns have raw scores of a 9, and the other 7 patterns have raw scores of 8.

We also calculated p-values based on the asymptotic distribution of standard errors. It results in only 10 patterns being rejected (see Table 23 in Appendix B). For the pattern with a raw score of 10, the MLE of the person parameter is not an interior point. So that pattern is excluded. Out of the 10 patterns with raw scores of 9, only 9 of them were rejected. It also rejects one pattern with a raw score of 8.

In this example, the exact test rejects more patterns compared to the asymptotic approach (Table 23 is a subset of Table 22). Some response patterns that are rejected under the exact test are failed to be rejected using the asymptotic approach. In practice, it would make a difference in decision making for those individuals with these response patterns.

2.6 Discussion

Being able to accurately recognize low reliabilities associated with short test forms is very important. Under the IRT framework, reliability is reflected as information or precision of the estimated latent abilities. Typically, the asymptotic distribution of the MLE is used to recognize the uncertainty of the person parameter estimates. However, the actual distribution is not normal even with 30 items for the 2PL model (Biehler et al., 2014). Developing a method that can obtain an accurate measure of the precision is of practical importance.

In this paper, we generalized the exact distribution approach for constructing the UMP test, equal-tail two sided test, and the associated confidence interval to IRT models within the exponential family. In addition, we proposed a branch and bound algorithm for the purpose of calculating the exact p-value.

Thissen (2016) argued that instead of reporting score in a yes-or-no fashion regarding a student's proficiency, we should report proficiency probabilistically. The method we proposed in this paper does *not* provide the probability of a student being proficient given a test score, which should be answered by using Bayesian posterior probabilities (Gelman et al., 2013). However, the exact p-value does provide evidence against the hypothesis that the student is proficient. The discussion of the interpretation of p-values dates back to Fisher (1935). Wasserman (2004) also gave the guidelines: $> .05$ being no evidence, $.01 - .05$ being positive evidence, .001-.01 being substantive evidence, and $< .001$ being decisive evidence against H_0.

The branch and bound algorithm introduced in this paper can be used to compute the exact p-value efficiently. If only the decision of whether to reject H_0 at the nominal level α is desired and the exact p-value is not of interest, the algorithm can be modified such that it terminates once the α level is reached. As a result, the algorithm could be even more efficient.

There has been some effort in exploring fast algorithms for the response pattern enumeration. One example is a network algorithm for computing person fit statistic under the Rasch model (Liou & Chang, 1992). The network algorithm uses a directed acyclic graph to represent response patterns. It deals with the enumeration of response patterns conditioned on a raw sum score. But the algorithm cannot handle the complete enumeration of all possi-

ble response patterns given a sufficient statistic constraint. The branch and bound algorithm developed in this paper tackles this problem by using a binary tree structure. Computing exact distributions of the person parameter shares a lot of similarities with computing exact distributions of the log-likelihood based person fit statistics (e.g. l_z statistic). In the latter case, there are two main differences. Instead of the hypothesized person parameter value, the computation of the person fit statistics should be conditioned on the estimated person parameter for the observed response pattern. Another major difference is the way how "extremity" of a pattern is defined. Compared to finding all patterns with greater (or smaller) weighted sum scores, calculating the exact distribution of the person fit statistics requires finding all patterns with worse fit (e.g. smaller log-likelihood). The branch and bound algorithm developed in this paper may be adapted for this task.

As mentioned, the introduced procedure for constructing the UMP test only works for IRT models within the exponential family where a sufficient statistic for the ability parameter exists and the monotone likelihood ratio property holds. It should be noted that the 3PL model is not in the exponential family. Thus, the method does not generalize to the 3PL model or any other model outside the exponential family. The same problem exists for the higher order approximation approach. The saddle-point approximation works only for the exponential family as well (Biehler et al., 2014). Developing appropriate inference procedures and efficient computational algorithms for the 3PL under small sample size still remains a challenge.

Appendix A

Under the 2PL model, the probability of a correct response for j^{th} item from a subject is

$$P_j(X_j = 1|a_j, b_j, \theta) = \frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]}, \tag{2.13}$$

where a_j is the item discrimination parameter, b_j is the item difficulty parameter, and θ is the ability parameter for the subject. It follows that the likelihood of θ given a response pattern $\boldsymbol{X} = \boldsymbol{x}$ is

$$L(\theta|\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}) = \prod_{j=1}^{J} P_j(X_j = 1|a_j, b_j, \theta)^{x_j} P_j(X_j = 0|a_j, b_j, \theta)^{1-x_j} \tag{2.14}$$

$$= \prod_{j=1}^{J} \{\frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]}\}^{x_j} \{\frac{1}{1 + \exp[a_j(\theta - b_j)]}\}^{1-x_j} \tag{2.15}$$

$$= \prod_{j=1}^{J} \frac{\{\exp[a_j(\theta - b_j)]\}^{x_j}}{1 + \exp[a_j(\theta - b_j)]} \{\frac{1}{1 + \exp[a_j(\theta - b_j)]}\}^{x_j - x_j} \tag{2.16}$$

$$= \prod_{j=1}^{J} \frac{\{\exp[a_j(\theta - b_j)]\}^{x_j}}{1 + \exp[a_j(\theta - b_j)]} \tag{2.17}$$

$$= \frac{\exp[\sum_{j=1}^{J} x_j a_j(\theta - b_j)]}{\prod_{j=1}^{J}\{1 + \exp[a_j(\theta - b_j)]\}} \tag{2.18}$$

$$= \frac{\exp[\theta \sum_{j=1}^{J} a_j x_j]}{\exp[\sum_{j=1}^{J} a_j x_j b_j] \prod_{j=1}^{J}\{1 + \exp[a_j(\theta - b_j)]\}} \tag{2.19}$$

$$= \exp[\theta \sum_{j=1}^{J} a_j x_j]\{\exp[\sum_{j=1}^{J} a_j x_j b_j]\}^{-1}\{\prod_{j=1}^{J}\{1 + \exp[a_j(\theta - b_j)]\}\}^{-1}. \tag{2.20}$$

The equation 2.20 is in the exponential form $L(\theta|\boldsymbol{x}) = \exp[\eta(\theta)T(\boldsymbol{x})]h(\boldsymbol{x})g(\theta)$, where

$$\exp[\eta(\theta)T(\boldsymbol{x})] = \exp(\theta \sum_{j=1}^{n} a_j x_j), \tag{2.21}$$

$$h(\boldsymbol{x}) = \exp(\sum_{j=1}^{n} a_j x_j b_j)^{-1}, \tag{2.22}$$

and

$$g(\theta) = \prod_{j=1}^{n}\{1 - \exp[a_j(\theta - b_j)]\}^{-1}. \tag{2.23}$$

Appendix B

In the food security data example, we are interested in testing the one-sided hypothesis: $H_0 : \theta \leq 1.93$ against $H_1 : \theta \geq 1.93$. Using the exact test approach, the following response patterns are rejected at $\alpha = 0.05$ level:

Table 22: 18 patterns that are rejected under the exact test

0	1	0	1	1	1	1	1	1	1
0	1	1	1	1	1	1	1	1	0
0	1	1	1	1	1	1	1	1	1
1	0	0	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1	1	1
1	1	0	0	1	1	1	1	1	1
1	1	0	1	0	1	1	1	1	1
1	1	0	1	1	1	1	1	1	0
1	1	0	1	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1	1
1	1	1	1	0	1	1	1	1	0
1	1	1	1	0	1	1	1	1	1
1	1	1	1	1	0	1	1	1	1
1	1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	1	1	0
1	1	1	1	1	1	1	1	1	1

Table 23: 10 patterns that are rejected under the asymptotic approach

0	1	1	1	1	1	1	1	1	1
1	0	1	1	1	1	1	1	1	1
1	1	0	1	1	1	1	1	1	0
1	1	0	1	1	1	1	1	1	1
1	1	1	0	1	1	1	1	1	1
1	1	1	1	0	1	1	1	1	1
1	1	1	1	1	0	1	1	1	1
1	1	1	1	1	1	0	1	1	1
1	1	1	1	1	1	1	0	1	1
1	1	1	1	1	1	1	1	0	1

Chapter 3

Estimating CDMs Using MCMC

3.1 Introduction

In the past two decades, Markov chain Monte Carlo (MCMC) techniques have been widely used for the Bayesian estimation of psychometric models. Not only an alternative to other estimation methods, MCMC algorithms and general purpose MCMC software has been facilitating the development of modern psychometric models that are otherwise difficult to fit (Levy, 2009). In this chapter, we provide a brief survey of MCMC methods used in estimating Cognitive Diagnostic Models (CDM). In addition, a Gibbs sampler for fitting the saturated Log-linear CDM model (LCDM, Henson, Templin, & Willse, 2009) is introduced. The utility of Bayesian inference is demonstrated by analyzing the Examination for the Certificate of Proficiency in English (ECPE) dataset.

To help understand the motivation of developing MCMC methods, consider the following general statistical inference problem. Given a set of observed data $\boldsymbol{X} = \boldsymbol{x}$, we would like to model the data with a probabilistic model $p(\boldsymbol{x}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the model parameter vector.

Under the Bayesian framework, a prior is assigned to the parameters, i.e. $p(\boldsymbol{\theta})$. Then we are interested in the posterior distribution of the model parameters given the observed data, i.e.

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{3.1}$$

In some cases the closed form of the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ can be analytically derived. However, under other circumstances, the posterior distribution must be approximated numerically. Difficulty arises from the numerical evaluation of the integral in the denominator of (3.1). If $\boldsymbol{\theta}$ is unidimensional, the integral can be approximated by using k quadrature points fairly efficiently. But in general, evaluating the multiple integral $\int\int\cdots\int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\theta_1 d\theta_2\cdots d\theta_d$ requires a high-dimensional grid of k^d points in $\mathbb{R}^d$. As the number of dimensions d grows, integration by quadrature quickly becomes infeasible. This problem is also referred to as the "curse of dimensionality". Instead of deterministically evaluating the high-dimensional integral, MCMC algorithms stochastically sample from the posterior distribution by constructing a Markov chain whose stationary distribution is the target posterior distribution. For a detailed review of MCMC, refer to Gelman et al. (2013), Neal (1998), and Brooks, Gelman, Jones, and Meng (2011).

Despite its importance to Bayesian inference, it should be noted that MCMC methods are not limited to Bayesian applications. High-dimensional integrals also arise from computing marginal maximum likelihood estimates in some models. As a result, MCMC as a class of efficient stochastic numerical integration algorithms is also used in frequentist applications. In fact, such applications have been developed in psychometrics. For example, Cai (2010) adapted the Metropolis-Hastings Robbins-Monro algorithm to estimate the high-dimensional item factor analysis model by marginal maximum likelihood. Given much

improved computing power and the availability of general purpose Bayesian inference software, the CDM literature, flourishing in relatively recent years, also saw a wide range of applications of MCMC methods.

3.2 MCMC Background

In this section, we provide a brief background and intuition of MCMC for readers who might not be familiar with the concept. A Markov chain is a *series* of random variables, $\boldsymbol{\Theta}^{(0)}, \boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Theta}^{(t+1)}, \ldots$, where the state at time $t+1$ depends only on the immediate previous state at t. In other words, the distribution of $\boldsymbol{\theta}^{(t+1)}$ is independent of everything else given $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\theta}^{(t)}$, i.e.

$$P(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(t)}) = P(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}). \tag{3.2}$$

This is often referred to as the Markov property. Additionally, the state space, that is the range of $\boldsymbol{\theta}$, is common across all time points. In practice, it implies the parameter space of the model cannot be changed. However, there exist MCMC methods that can handle models with variable parameter space - for example, the reversible jump MCMC. This topic is significantly more advanced and out of the scope of this chapter. Interested readers may refer to P. J. Green (1995). Observing the aforementioned Markov property, it is clear that, in order to define a Markov chain, we need to specify the probability of an initial state $\boldsymbol{\theta}$ - $p_0(\boldsymbol{\theta}) = P(\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta})$ and the transition probabilities between consecutive states - $T_t(\boldsymbol{\theta}, \boldsymbol{\theta}') = P(\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta})$ for $t = 0, 1, 2, \ldots$. Then the distribution of $\boldsymbol{\theta}$ at time

$t+1$ can be determined by

$$p_{t+1}(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{\theta}}} p_t(\tilde{\boldsymbol{\theta}}) T_t(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}). \tag{3.3}$$

For *homogeneous* Markov chains, the transition probabilities stay the same across all time points, i.e. $T_t(\boldsymbol{\theta}, \boldsymbol{\theta}') = T(\boldsymbol{\theta}, \boldsymbol{\theta}'), \forall t$. A Markov chain is said to have reached its *stationary* or *invariant* distribution - $\pi(\boldsymbol{\theta})$ if the distribution of $\boldsymbol{\theta}$ does not change according to time points t any more. Specifically, there exists some $\tilde{t}$ such that $p_{\tilde{t}}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ and

$$\pi(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{\theta}}} \pi(\boldsymbol{\theta}) T_t(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}), \forall t \geq \tilde{t}. \tag{3.4}$$

The purpose of using MCMC in Bayesian inference is to help us sample from an otherwise difficult posterior distribution. To achieve this goal, we are interested in constructing a Markov chain where the target posterior distribution is invariant. Often, we choose reversible homogeneous Markov chains in which the probability of a transition from the state $\boldsymbol{\theta}$ to the state $\boldsymbol{\theta}'$ is the same as the probability of a transition from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}$ under the distribution of states π. Equivalently,

$$\pi(\boldsymbol{\theta}) T(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}') T(\boldsymbol{\theta}', \boldsymbol{\theta}). \tag{3.5}$$

The above condition is usually called *detailed balance*. It is straightforward to show that detailed balance implies invariance, i.e.

$$\sum_{\boldsymbol{\theta}'} \pi(\boldsymbol{\theta}') T(\boldsymbol{\theta}', \boldsymbol{\theta}) = \sum_{\boldsymbol{\theta}'} \pi(\boldsymbol{\theta}) T(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}) \sum_{\boldsymbol{\theta}'} T(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}). \tag{3.6}$$

It should be noted that detailed balance is a sufficent but not necessary condition for a distribution to be invariant (Neal, 1998).

Detailed balance ensures that once a Markov chain reaches its invariant distribution, subsequent states are samples from this invariant distribution. However, we generally do

not know this invariant distribution which is the target posterior distribution. Instead, we hope the distribution of states at time t converges in distribution to its invariant distribution π as $t \to \infty$ regardless of its initial probability distribution of states $p_0(\boldsymbol{\theta})$. The Markov chain is *ergodic* if it holds this property. For a homogeneous Markov chain with an invariant distribution π, it is ergodic if the chain can traverse the entire support of π, i.e.

$$v = \min_{\boldsymbol{\theta}} \min_{\boldsymbol{\theta}':\pi(\boldsymbol{\theta}')>0} T(\boldsymbol{\theta}, \boldsymbol{\theta}')/\pi(\boldsymbol{\theta}') > 0. \tag{3.7}$$

For a proof of this theorem, readers can refer to Neal (1998).

The simplest MCMC algorithm is perhaps the Gibbs sampler (Geman & Geman, 1984; Gelfand & Smith, 1990). Suppose we are interested in sampling from a joint distribution given by $p(\theta_1, \theta_2, \ldots, \theta_K)$ which is our target posterior distribution. Gibbs sampler works by repeatedly sampling each θ_k from their full conditional distributions. At the t^{th} iteration, we

- sample $\theta_1^{(t)}$ according to the distribution given by $p(\theta_1^{(t)}|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_K^{(t-1)})$;
- sample $\theta_2^{(t)}$ according to the distribution given by $p(\theta_2^{(t)}|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_K^{(t-1)})$;

$$\vdots$$

- sample $\theta_K^{(t)}$ according to the distribution given by $p(\theta_K^{(t)}|\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{K-1}^{(t)})$.

The above steps together form a transition of state from $\boldsymbol{\theta}^{t-1}$ to $\boldsymbol{\theta}^t$ with probabilities $T(\boldsymbol{\theta}, \boldsymbol{\theta}')$ that leaves the target distribution invariant. Starting from an initial state $\boldsymbol{\theta}^{(0)}$, after simulating the Markov chain long enough, subsequent draws of $\boldsymbol{\theta}^{(t)}$ are treated as samples from the target posterior distribution.

3.3 Applications of MCMC in CDM

Similar to other types of psychometric modeling, instances of applications of MCMC in CDM are numerous. By no means the brief survey in this section is exhaustive, but rather to give readers flavors of the existing literature. The applications of MCMC in CDM can be traced back to earlier papers on the topic. In Junker and Sijtsma (2001), one of the earlier papers on CDM, the authors fit the deterministic inputs, noisy "and" gate (DINA) model and the noisy inputs, deterministic "and" gate (NIDA) model using the BUGS (Bayesian inference Using Gibbs Sampling) software (Thomas, Spiegelhalter, & Gilks, 1992).

While de la Torre (2008) provides a reference work for estimating the DINA model by marginal maximum likelihood using the expectation-maximization (EM; Dempster, Laird, & Rubin, 1977) algorithm; the development of the EM algorithm for the higher-order DINA (HO-DINA) model is not trivial. As a result, in de la Torre and Douglas (2004), the HO-DINA model is estimated by a blocked Gibbs sampler (Geman & Geman, 1984; Gelfand & Smith, 1990; Gelman et al., 2013). The full-conditional distributions for HO-DINA do not have closed forms and are not easy to sample from directly. Therefore, the authors adopted the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970). Instead of directly sampling from the full-conditional distributions, at each iteration, the Metropolis algorithm draws a proposed sample value or vector from a proposal distribution (usually a Gaussian distribution), and accepts or rejects it with an appropriately defined acceptance probability. To calculate the acceptance probability, only the unnormalized full conditional density function is required. This circumvents the difficulty of obtaining the normalizing constant when it cannot be derived analytically. The combination

of Gibbs sampler and Metropolis algorithm is usually referred to as the Metropolis-within-Gibbs which is implemented in many general purpose Bayesian inference softwares (e.g. OpenBUGS;Lunn, Spiegelhalter, Thomas, & Best, 2009, JAGS;Plummer, 2005).

One difficulty of using the Metropolis algorithm is the tuning of the sampler. If the variance of the proposal distribution is large, the proposed sample is more likely to be further away from the current sample, which leads to low acceptance probabilities. Consequently, a large number of proposed samples are rejected before an acceptance, and the sampler rarely moves. On the other hand, small variance of the proposal distribution leads to high acceptance probabilities. But the proposed samples tend to be close to the current ones. As a result, the sampler moves slowly and does not explore the posterior distribution very efficiently. Therefore tuning is required so that the Markov chain is mixing at an optimal rate (Roberts, Gelman, & Gilks, 1997). Tuning a sampler could be a tedious task. Culpepper (2015) derived the closed forms of full-conditional distributions for DINA model so that the parameters can be directly sampled without using the Metropolis algorithm. In the same paper, the author also shows that the monotonicity assumption of the DINA model can be enforced by sampling the item parameters from a truncated bi-variate Beta distribution.

In the applications discussed so far, the Q-matrix (Tatsuoka, 1983) needs to be specified before the model can be estimated. In reality, the specification of the Q-matrix is not always straightforward and elements of the Q-matrix can be uncertain. Recognizing this limitation, DeCarlo (2012) proposed a Bayesian model to handle the uncertainty. Instead of treating all elements of the Q-matrix as fixed, the author specifies some of them as Bernoulli distributed random parameters, and assigns a Beta prior to the Bernoulli probabilities. The uncertain elements of the Q-matrix are recovered from examining the posterior distributions. Open-

Bugs software (Spiegelhalter, Thomas, Best, & Lunn, 2014) is used to estimate the model under the reparameterized DINA (RDINA;DeCarlo, 2012) model. DeCarlo and Kinghorn (2016) extend the approach to the case where none of the Q-matrix elements is fixed.

Furthermore, there has also been some other effort developing exploratory Bayesian methods for estimating CDM models without any prior knowledge of the Q-matrix except for the dimensions. Chung (2014) derives a Gibbs sampler for the DINA model and a Metropolis-within-Gibbs algorithm for the rRUM(reduced reparameterized unified model; Hartz, 2002) that include all elements of the Q-matrix as model parameters. The distribution of the attribute patterns for examinees is modeled by a saturated categorical distribution, and the probabilities of the categories are given a Dirichlet prior. Thanks to the categorical-Dirichlet conjugacy, the probabilities of attribute patterns can be directly sampled from Dirichlet posterior distributions. By using a saturated categorical distribution, the author did not assume a particular factorization of the joint distribution of the attributes. Correlated attributes with different structures can be modeled in addition to independent attributes. However, the trade-off is the large number of parameters needs to be estimated. For a Q-matrix with K attributes, there are $2^K - 1$ probabilities for the attribute patterns. The Q-matrix is estimated similarly by using a categorical distribution. Item parameters for the DINA model can be sampled from truncated Beta distributions respecting the monotonicity assumption. Unfortunately, the full-conditional distributions of the item parameters for the rRUM model do not have closed forms. Thus, the Metropolis algorithm is used. Another example of the exploratory Bayesian approach can be found in Chen, Culpepper, Chen, and Douglas (2018). The paper deals with the same problem of estimating the DINA model without knowing the elements of the Q-matrix. Building on the development in understanding the identifiablity

of the DINA model (Chen, Liu, Xu, & Ying, 2015; J. Liu, Xu, & Ying, 2012, 2013; Xu & Zhang, 2016), Chen et al. (2018) constrain the Q-matrix to be identified in their estimation procedure.

MCMC also aids the development and applications of more complex CDM models. For example, Li, Cohen, Bottge, and Templin (2016) introduce a longitudinal model that incorporates learning into CDM models. The attribute patterns for each student can change over time. It is modeled by a latent transition model. The transition matrix indicates the probability of transition from one attribute pattern to another. In this paper, several models with different transition matrices are fitted and compared using deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002).

Not only useful in estimating CDM models, MCMC also provides some of the most intuitive ways in checking model fit. Using the posterior samples, the posterior predictive model check (PPMC) method (Rubin, 1984; Gelman et al., 2013) calculates posterior distributions of various fit measures. It has been used in assessing the fit of IRT models (Sinharay, 2005). In CDM, Park, Johnson, and Lee (2015) examines the performance of PPMC using observed total-scores distribution, association of item pairs, and correlation of attribute pairs in assessing model fit.

As we mentioned earlier, the review in this section is far from exhaustive. As more and more elaborate CDM models are developed in literature, we will certainly see more applications of MCMC.

3.4 A Gibbs sampler for the saturated log-linear CDM model

In this section, we propose a new Gibbs sampler for the LCDM model. We analyze the ECPE data set as an illustration.

3.4.1 The log-linear CDM model

The LCDM is similar to the Generalized DINA(GDINA;de la Torre, 2011) model in the sense that they all provide a flexible and general framework that encompasses many specific CDM models and can be viewed as a special case of the general diagnostic model (GDM; von Davier, 2008, 2014).

Under the LCDM, the probability of nth person answering the kth item correctly is

$$\text{logit}\, P_k(\boldsymbol{a}_n) = \lambda_{k0} + \sum_{d=1}^{K} \lambda_{kd} a_{nd} q_{kd} + \sum_{d=1}^{K} \sum_{d'>d} \lambda_{kdd'} a_{nd} a_{nd'} q_{kd} q_{kd'} + \dots . \tag{3.8}$$

$a_{nd} \in \{0, 1\}$ with $a_{nd} = 1$ being the nth person has the dth attribute, and $a_{nd} = 0$ otherwise. Similarly, q_{kd} is 1 if the kth item measures the dth attribute, and 0 otherwise. λ_{k0} is the intercept, so a person does possess any of the skills measured in the test would have the probability of $\text{logit}^{-1}(\lambda_{k0})$ getting the kth item correct. λ_{kd} is the main effect for the dth attribute. And $\lambda_{kdd'}$ is the interaction effect for the dth and d'th attributes. Depending on the Q-matrix, some of the terms in (3.8) may be dropped. If an item only measures one attribute, there is only the intercept and one main effect. It should be noticed that some specific CDM models are nested within (3.8). For example, if only the highest order interaction and the intercept are retained, the LCDM reduces to the DINA. A saturated

model includes the intercept, all main effects of the measured attributes, and all interaction terms associated with those attributes.

3.4.2 A Bayesian Formulation of the Reparameterized Saturated LCDM

For a general CDM with three attributes, there can be $2^3 = 8$ latent classes defined by the attribute patterns $\boldsymbol{a}$; therefore, under the unrestricted latent class model, there would be 8 item response probabilities that would need to be estimated for each item. The Q-matrix restricts the probabilities by enforcing certain equality constraints on the item response probabilities. For example, under the saturated LCDM, the probability of giving a correct response to an item by different people who possess different subsets of the required attributes may be different. Suppose an item requires the first two attributes but not the third, so the kth row of the Q-matrix is $\boldsymbol{q}_k = (1, 1, 0)$. Then three people with attribute patterns $\boldsymbol{a} = (1, 1, 0)$, $(1, 0, 0)$, and $(0, 1, 0)$ may potentially receive different probabilities of giving a correct answer to this item. However, a person with the attribute pattern $(1, 1, 1)$ would have the same probability of giving a correct response as someone whose attribute pattern is $(1, 1, 0)$ due to the fact that the third attribute is not required by the item. As a result, there are $2^2 = 4$ probabilities associated with this item. Except for this restriction, the saturated model LCDM does not make any further constraints.

In the following Bayesian specification of the item-saturated LCDM, we use the natural probabilities as the model parameters rather than using the linear coefficients. To aid in the description of the model, we define the condensed attribute pattern $\boldsymbol{\omega}_{nk}$ for each individual

n and item k, as the subvector of $\boldsymbol{a}_n$ corresponding to only the dimensions or attributes required by item k, i.e., $\boldsymbol{\omega}_{nk} = (\boldsymbol{e}_{d_1}, \ldots, \boldsymbol{e}_{d_m})^\top \boldsymbol{a}_n$, where $\boldsymbol{e}_d$ is the standard unit vector for dimension d with a 1 for element d and a zero everywhere else, and the $\boldsymbol{d}$ is an ordered index set $\boldsymbol{d} = \{m : q_{km} = 1\}$. In our three attribute example, with only the first two attributes required for an item, we have

$$\boldsymbol{\omega}_{nk} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \boldsymbol{a}_n.$$

Then the item response probability can be denoted by $p_{k(\boldsymbol{\omega}_{nk})} = P(X_{nk} = 1 | \boldsymbol{A} = \boldsymbol{a})$.

Formally, suppose we observe an N by K response matrix from N subjects answering K items and a K by D Q-matrix, then our Bayesian hierarchical formulation of the LCDM assumes

$$x_{nk} | \boldsymbol{\omega}_{nk}, \boldsymbol{p}_k \sim \text{Bernoulli}(p_{k(\boldsymbol{\omega}_{nk})}), \tag{3.9}$$

$$p_{k(\boldsymbol{\omega}_{nk})} \sim \text{Beta}(\alpha_k, \beta_k), \tag{3.10}$$

$$\boldsymbol{\alpha}_n | \boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi}), \tag{3.11}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{v}). \tag{3.12}$$

Conditional on the latent attributes required by a particular item $\boldsymbol{\omega}_{nk}$, a person gives a correct response with the probability $p_{k(\boldsymbol{\omega}_{nk})}$. We assume a Beta prior distribution for the vector of item response probabilities $\boldsymbol{p}_k$. A non-informative prior can be specified by giving the uniform Beta$(1, 1)$; while, a Beta(0.5,0.5) may be used if a researcher believes the item might have a higher discrimination among those with and without the required skills.

We do not assume a particular factorization of the joint distribution of the attributes. Instead, each of the possible 2^D attribute patterns is treated as a category. Then each

person's attribute pattern follows a categorical distribution with probabilities of each possible attribute pattern governed by parameters $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{2^D})^\top$. A Dirichlet hyper-prior with concentration parameters $\boldsymbol{v}$ is given to the categorical distribution parameters.

3.4.3 Monotonicity Constraint

The monotonicity assumption specifies a set of constraints that ensures the interpretability of CDM models in addition to the specification of the Q-matrix. Under the monotonicity assumption, mastering additional attributes would not lower the probability of giving a correct response, i.e.

$$P(X_{nk} = 1|\boldsymbol{a}_{n_1}) \geq P(X_{nk} = 1|\boldsymbol{a}_{n_2}), \tag{3.13}$$

whenever $\boldsymbol{\omega}_{n_1kd} \geq \boldsymbol{\omega}_{n_2kd}$ for all $d = 1, \ldots, D_k$, where D_k is the number of skills required by item k. Thus the item parameters in our Bayesian hierarchical formulation must satisfy

$$p_{k(\boldsymbol{\omega}_{n_1k})} \geq p_{k(\boldsymbol{\omega}_{n_2k})}, \text{if } \omega_{n_1kd} = 1 \ \forall \ d \text{ s.t. } \omega_{n_2kd} = 1. \tag{3.14}$$

For the log-linear model, it is equivalent to constraining all main effects to be nonnegative and the coefficient of any interaction term to be no less than -1 times the largest main effect involved in the interaction (Henson et al., 2009; Templin & Bradshaw, 2014).

3.4.4 A Gibbs Sampler

Conditional on the observed data for the kth item and class assignment for all people on this item, the item parameter is independent of everything else. So its full conditional

distribution is

$$P(p_{k(\boldsymbol{w})}|\boldsymbol{x}_{(k)},\boldsymbol{\omega}_{(k)},\alpha_k,\beta_k) \propto \prod_{S_{k\boldsymbol{w}}=\{n:\ \boldsymbol{\omega}_{nk}=\boldsymbol{w}\}} p_{k(\boldsymbol{w})}^{x_{nk}}(1-p_{k(\boldsymbol{w})})^{1-x_{nk}} P(p_{k(\boldsymbol{w})}|\alpha_k,\beta_k), \tag{3.15}$$

where $\boldsymbol{x}_{(k)}$ denotes the vector of all item responses to item k and $\boldsymbol{\omega}_{(k)}$ denotes the set of item-specific attribute patterns for item k.

Due to the standard Bernoulli-Beta conjugacy, (3.15) has a closed form, i.e.

$$p_{k(\boldsymbol{w})}|\boldsymbol{x}_{(k)},\boldsymbol{\omega}_{(k)},\alpha_k,\beta_k \sim \text{Beta}\left(\alpha_k + \sum_{n\in S_{k\boldsymbol{w}}} x_n, \beta_k + |S_{k\boldsymbol{w}}| - \sum_{n\in S_{k\boldsymbol{w}}} x_n\right). \tag{3.16}$$

The monotonicity constraint in (3.14) implies that $p_{j(\boldsymbol{\omega}_{ij})}$ is bounded above by

$$U_{p_{k(\boldsymbol{w})}} = \inf_{\boldsymbol{w}'}\{p_{k(\boldsymbol{w}')} :\ w'_d \geq w_d\ \forall\ d \in \{1,2,\ldots,D_k\}\}, \tag{3.17}$$

and bounded below by

$$L_{p_{k(\boldsymbol{w})}} = \sup_{\boldsymbol{w}'}\{p_{k(\boldsymbol{w}')} :\ w'_d \leq w_d\ \forall\ d \in \{1,2,\ldots,D_k\}\}. \tag{3.18}$$

It follows that the full conditional distribution in (3.16) should be truncated, i.e.

$$p_{k(\boldsymbol{w})}|\boldsymbol{x}_{(k)},\boldsymbol{\omega}_{(k)},\alpha_k,\beta_k \sim \text{Beta}\left(\alpha_k + \sum_{n\in S_{k\boldsymbol{w}}} x_{nk},\ \beta_k + |S_{k\boldsymbol{w}}| - \sum_{n\in S_{k\boldsymbol{w}}} x_n\right) I_{(L_{p_{k(\boldsymbol{w})}},U_{p_{k(\boldsymbol{w})}})}(p_{k\boldsymbol{w}}), \tag{3.19}$$

where $I_{(u,\ell)}(p)$ indicates the distribution is truncated to the interval (u,ℓ).

The full conditional distribution for $\boldsymbol{a}_n$ is

$$P(\boldsymbol{a}_n|\boldsymbol{x}_n,\boldsymbol{p},\boldsymbol{\pi}) \propto \prod_{k=1}^{K} P(x_{nk}|\boldsymbol{p}_{k(\boldsymbol{\omega}_{nk})})P(\boldsymbol{a}_n|\boldsymbol{\pi}). \tag{3.20}$$

Since the distribution is discrete, (3.20) can be easily normalized:

$$P(\boldsymbol{a}_n|\boldsymbol{x}_n,\boldsymbol{p},\boldsymbol{\pi}) = \frac{\prod_{k=1}^{K} P(x_{nk}|\boldsymbol{p}_{k(\boldsymbol{\omega}_{nk})})P(\boldsymbol{a}_n|\boldsymbol{\pi})}{\sum_{\boldsymbol{a}_n}\prod_{k=1}^{K} P(x_{nk}|\boldsymbol{p}_{k(\boldsymbol{\omega}_{nk})})P(\boldsymbol{a}_n|\boldsymbol{\pi})}. \tag{3.21}$$

And the closed form full-conditional distribution is

$$\boldsymbol{a}_n|\boldsymbol{x}_n, \boldsymbol{p}, \boldsymbol{\pi} \sim \text{Categorical}(u_1, u_2, \ldots, u_{2^D}), \tag{3.22}$$

where the probabilities $u_1, u_2, \ldots, u_{2^D}$ are given in (3.21).

Finally, the full conditional distribution for hyper-parameters $\boldsymbol{\pi}$ is

$$P(\boldsymbol{\pi}|\boldsymbol{a}, \boldsymbol{v}) \propto \prod_{n=1}^{N} P(\boldsymbol{a}_n|\boldsymbol{\pi})P(\boldsymbol{\pi}|\boldsymbol{v}). \tag{3.23}$$

The standard categorical-Dirichlet conjugacy leads to the closed form:

$$\boldsymbol{\pi}|\boldsymbol{a}, \boldsymbol{v} \sim \text{Dirichlet}(\boldsymbol{v} + (c_1, c_2, \ldots, c_{2^d})), \tag{3.24}$$

where the elements of the vector $(c_1, c_2, \ldots, c_{2^D})$ are the counts of observations in each class.

Update steps for each iteration of the Gibbs sampler are:

1. Draw the item parameters $p_{k\boldsymbol{w}}$ for each item and item-specific attribute pattern $\boldsymbol{w}$ from the full conditional distributions in (3.19);

2. Draw the the latent class assignment $\boldsymbol{a}_n$ for each person from the full conditional distributions in (3.22);

3. Draw the hyper-parameter $\boldsymbol{\pi}$ from the full conditional distribution in (3.24).

3.4.5 Linear Transformation of Model Parameters

The model parameters from the reparameterized saturated model can be easily transformed back to the log-linear model parameters by solving a linear system of equations. For simplicity, consider the case where there are $d = 2$ attributes. Under the saturated log-linear

model, $2^2 = 4$ linear coefficients are needed. The logit link links the probabilities to the linear combinations of the attributes, i.e.

$$\boldsymbol{T}\boldsymbol{\lambda}_k = \text{logit}\,\boldsymbol{p}_k, \tag{3.25}$$

where

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \boldsymbol{\lambda}_j = \begin{bmatrix} \lambda_{k0} \\ \lambda_{k1} \\ \lambda_{k2} \\ \lambda_{k12} \end{bmatrix}, \text{logit}\,\boldsymbol{p}_k = \begin{bmatrix} \text{logit}\,p_{k(00)} \\ \text{logit}\,p_{k(10)} \\ \text{logit}\,p_{k(01)} \\ \text{logit}\,p_{k(11)} \end{bmatrix}.$$

In the above notations, $\boldsymbol{t}_{m\cdot}$ denotes the mth row of the T matrix. Multiplying the inverse of the attribute pattern matrix to both sides of (3.25) gives the log-linear model parameters, i.e.

$$\boldsymbol{\lambda}_k = \boldsymbol{T}^{-1}\,\text{logit}\,\boldsymbol{p}_k. \tag{3.26}$$

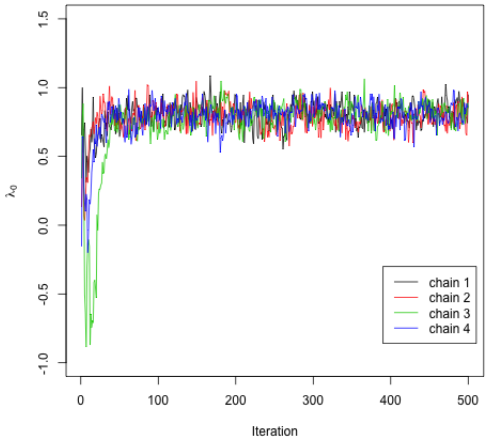
To get the posterior distribution of the log-linear model parameters, simply apply the linear transformation in (3.26) to the posterior samples of the reparameterized saturated model parameters.

3.5 A Bayesian Analysis of the ECPE Dataset

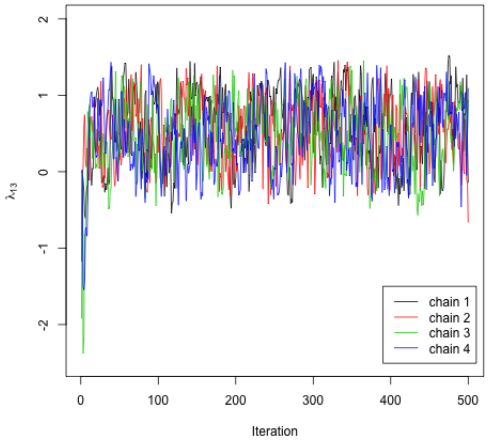
In this section, we analyze the ECPE dataset as a demonstration. The ECPE dataset is available in the *R* CDM package (George, Robitzsch, Kiefer, Groß, & Ünlü, 2016). It has been analyzed in previous research (e.g. Templin & Bradshaw, 2014; Templin & Hoffman, 2013). The dataset consists of the binary responses from 2922 examinees to 28 items. Three attributes are specified in the Q-matrix: morphosyntactic rules, cohesive rules, and

lexical rules. However, none of the items measures all three attributes. Among the 28 items, 9 measure two attributes, and the rest measure one. We fit the reparameterized saturated model and finally transformed parameters back to the log-linear model parameterization. Non-informative priors are used: uniform Beta$(1, 1)$ for item parameters, and Dirichlet$(1, 1, \ldots, 1)$ for the hyper-prior of class allocations. Futhermore, the monotonicity is enforced by imposing constraints to item parameters as in (3.17) and (3.18).

Diagnosing the convergence of the Markov chains is important in applications of MCMC. The MCMC theory guarantees that the Gibbs sampler will eventually converge to the target posterior distribution as the number of draws goes to infinity. But, in reality, the number of draws we can afford is always finite and often limited. Therefore, we need to assess whether we can treat MCMC draws approximately as samples from the posterior distribution after a certain number of initial draws. Over the years, many MCMC convergence diagnostics have been proposed. Some of the popular examples include the potential scale reduction factor (PSRF;Gelman & Rubin, 1992), the multivariate PSRF(MPSRF;Brooks & Gelman, 1998), and the Geweke convergence diagnostic (Geweke, 1992). Here we use two common graphical methods to assess the convergence of our Gibbs sampler. Four parallel chains with different starting values are simulated. We run each chain for 5000 iterations. To demonstrate the evidence of convergence, Figure 31 shows the trace of the first 500 iterations of each chain for two parameters. The plots suggest that the chains quickly converged to their target stationary distributions regardless of different starting values. We can also monitor the convergence by examining the k-lag autocorrelation functions. The k-lag autocorrelation is the correlation between every draw and its kth lag. Intuitively, a Markov chain that generates highly correlated samples would take a long time to explore the entire target distribution.

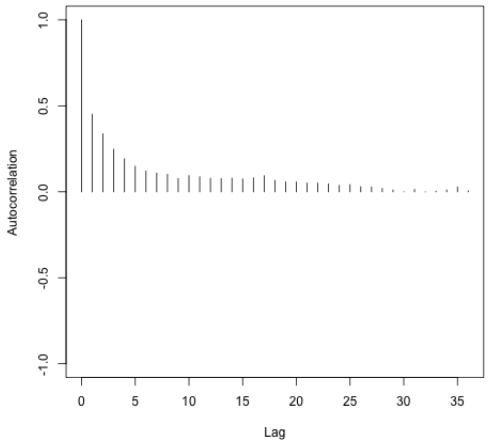


(a) λ_0 - Item 1

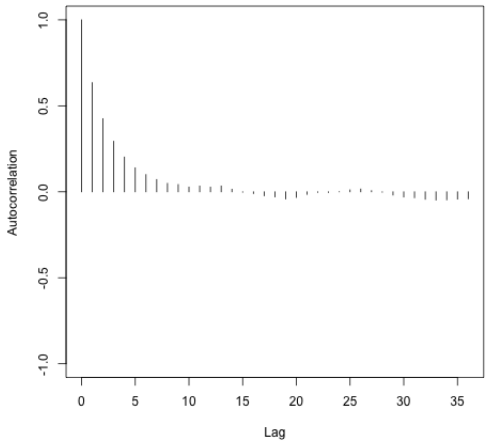


(b) λ_{13} - Item 11

Figure 31: k-lag autocorrelation of two parameters



(a) λ_0 - Item 1



(b) λ_{13} - Item 11

Figure 32: Traceplot of two parameters

We would hope that the autocorrelation between samples quickly shrink to around zero as the lag k increases. From Figure 32, we can see that the autocorrelation decreases very quickly as the lag increases in both cases. It is consistent with the quick convergence and good mixing shown in the trace plots. Based on the convergence diagnostics, our Gibbs sampler seems to perform very well. We decide to treat the first 1000 from each chain as

burn-ins and use the rest for the purpose of posterior inference.

Table 31 shows the *Expected a Priori* (EAP) estimates and posterior standard deviations of item parameters under the LCDM. Comparing the EAP estimates to the maximum likelihood estimates (MLE) reported in previous literature (see Table 1 in Templin & Bradshaw, 2014), it seems that the EAP estimates are almost identical to the MLE for the items measuring single attributes. However, differences exist between the EAP estimates and the MLE for items measuring two attributes except for the second item where the Bayesian approach yields similar estimates to maximum likelihood.

As pointed out by Templin and Bradshaw (2014), a closer examination of the MLE for two attribute items reveals that many of the ML estimates appeared on the boundary. For example, the main effect of the morphosyntactic rules for the first item is estimated to be zero in Templin and Bradshaw (2014). The standard asymptotic theory does not give any useful approximation to the limiting distribution of the MLE when the ML estimate lies on the boundary. This is reflected by the zero standard error reported in Templin and Bradshaw (2014). The MLE for some of the interaction effects also suffer this problem. They are estimated to be very close to the boundary imposed by the monotonicity constraint. Large standard errors are also observed for many of the estimated effects. These are symptoms of under-identification. von Davier (2014) also discussed this problem. While an infinitely large sample size would allow the parameters to be estimated precisely and away from boundaries (when the true parameters are away from the boundary), we work with a limited sample size in reality.

One solution is to impose an attribute hierarchy which effectively reduces the number of parameters to be estimated (Templin & Bradshaw, 2014). The introduced Bayesian method

Table 31: ECPE Bayesian estimates of LCDM item parameters

item	λ_0	λ_1	λ_2	λ_3	λ_{12}	λ_{13}	λ_{23}
1	0.81(0.08)	0.51(0.4)	0.65(0.23)		0.61(0.53)		
2	1.03(0.08)		1.25(0.15)				
3	-0.34(0.08)	0.76(0.42)		0.35(0.13)		0.52(0.44)	
4	-0.14(0.08)			1.69(0.1)			
5	1.07(0.08)			2.02(0.16)			
6	0.87(0.08)			1.68(0.14)			
7	-0.09(0.08)	1.59(0.67)		0.93(0.13)		0.32(0.7)	
8	1.47(0.09)		1.92(0.24)				
9	0.12(0.07)			1.19(0.1)			
10	0.05(0.06)	2.05(0.15)					
11	-0.05(0.08)	1.19(0.6)		0.96(0.14)		0.39(0.64)	
12	-1.79(0.12)	0.62(0.46)		1.31(0.17)		0.88(0.49)	
13	0.66(0.06)	1.61(0.15)					
14	0.17(0.05)	1.36(0.12)					
15	0.99(0.08)			2.12(0.16)			
16	-0.09(0.08)	1.34(0.57)		0.87(0.13)		0.13(0.59)	
17	1.34(0.09)		0.65(0.41)	0.61(0.27)			0.2(0.52)
18	0.92(0.08)			1.4(0.13)			
19	-0.2(0.08)			1.85(0.11)			
20	-1.43(0.1)	0.97(0.58)		0.94(0.15)		0.67(0.61)	
21	0.16(0.08)	0.98(0.54)		1.13(0.14)		0.12(0.58)	
22	-0.87(0.09)			2.24(0.11)			
23	0.66(0.08)		2.06(0.19)				
24	-0.69(0.09)		1.54(0.12)				
25	0.09(0.05)	1.14(0.11)					
26	0.16(0.08)			1.12(0.1)			
27	-0.89(0.06)	1.7(0.1)					
28	0.56(0.08)			1.75(0.12)			

Note: Attribute 1 - Morphosyntactic rules; Attribute 2 - Cohesive rules; Attribute 3 - Lexical rules.

explores another approach. The use of priors provides regularization and enables more parameters to be reasonably estimated (Gelman et al., 2013). The EAP estimates for single attribute items are well-regularized with small posterior standard deviations. While the posterior standard deviations for the two attribute items are larger, they are still reasonable. The largest posterior standard deviation is 0.67 compared to the largest standard error of 1.62 reported in previous research.

The posterior samples can also provide useful information in assessing various aspects of model fit. For example, one source of the misfit is the misspecification of the Q-matrix. Considering the EAP estimates and the associated posterior standard deviations of λ_1 and λ_{13} for item 20 in Table 31, one might suspect that morphosyntactic rules are not measured by the item. Both marginal posterior distributions of λ_1 and λ_{13} might have considerable densities around zero. However, if the item doesn't measure this attribute, it would imply that $\lambda_1 = \lambda_{13} = 0$. In other words, we need to inspect the joint posterior distribution of these two effects. Samples from the posterior simulation can achieve this with little effort. Figure 33 clearly shows that the origin is away from the region where the joint posterior density is concentrated.

Posterior samples can also be used to check the plausibility of particular CDM models. For example, if the DINA is plausible, it would suggest that the main effects and lower order interactions are all zeros. Since each item measures at most two attributes in the ECPE dataset, we only need to examine the joint posterior distribution of the main effects. Figure 34 suggests that DINA is more plausible for Item 1 than Item 11.

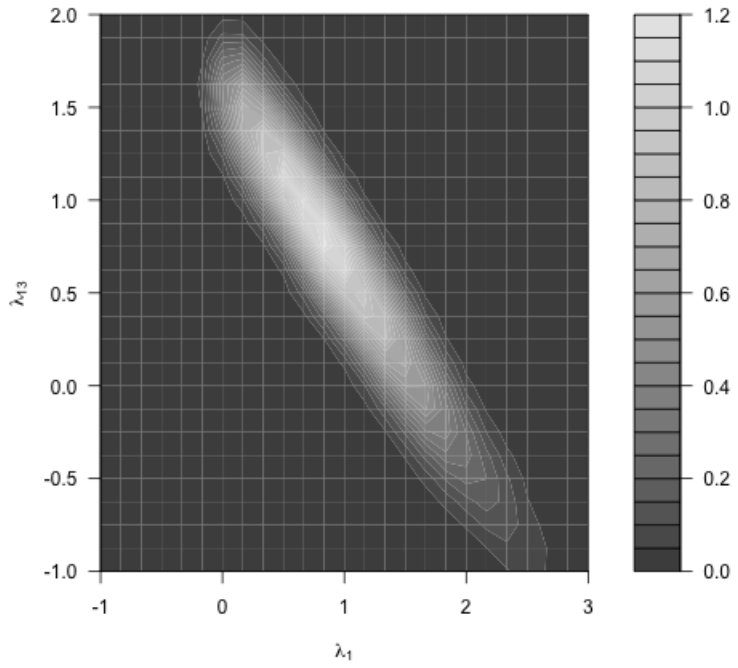
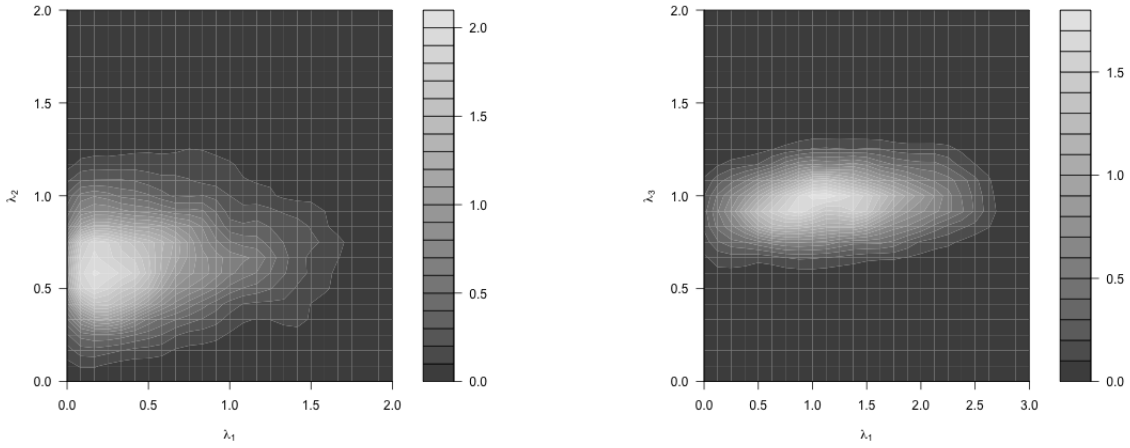


Figure 33: Joint posterior density of λ_1 and λ_{13} for Item 20



(a) Item 1

(b) Item 11

Figure 34: Joint posterior density of the main effects for Items 1 and 11.

3.6 Discussion

MCMC algorithms and Bayesian methods in general will certainly continue to play an important role in the development of various CDM models. In this chapter, we briefly reviewed

some of the applications of the MCMC in CDM literature. We also introduced a Gibbs sampler for estimating the saturated LCDM model. With the reparameterization, the sampler is able to take advantage of the standard conjugacy results thus the sampler does not require any tuning. Even though we introduced the sampler for the saturated LCDM, the approach can be modified to fit a wide spectrum of specific CDM models by imposing additional constraints to the saturated LCDM model.

Chapter 4

Bayesian Ordered Latent Class Models

4.1 Introduction

Many psychometric models have been developed to scale response data which is a common task in educational and psychological measurement. Item response theory (IRT; Hambleton & Swaminathan, 1985) models assume the items are measuring a continuous latent trait. Latent class models (LCM; B. Green, 1951; Goodman, 1974; Skrondal & Rabe-Hesketh, 2007), on the other hand, deal with the categorical latent trait. The standard LCM treats the latent classes as nominal. In other words, there is no stochastic ordering property among the latent classes. This is in contrast to the monotonicity assumption in IRT models where the probability of a correct response increases as the latent trait increases. Even though LCM is routinely used for the classification multivariate data, there are cases in which the stochastic ordering assumption is important in interpreting the latent classes. Croon (1990)

introduced the ordered latent class model (OLCM) by imposing the stochastic ordering constraints on the item parameters. Some recent examples of OLCM include ranking medical procedures in terms of the distribution of patient morbidity following the procedures (Yang, O'Brien, & Dunson, 2011), examining gender and country differences in TIMSS through the regression extension of OLCM (Cha, 2011), and analyzing polytomous questionnaire items about coping strategies with industrial malodor (van Onna, 2002).

One difficulty in applying LCM and OLCM is to select the number of classes K. In some applications, K is specified *a priori* based on some prior knowledge. However, in other cases, prior knowledge of K is not available which requires K to be determined in an exploratory fashion. Generally, there are two classes of exploratory approaches in determining K. The first class treats K as a modeling choice. Thus, it has to be specified before the model can be fitted. To select the best number of classes, models with different Ks are fitted. Then the selection of K is based on finding the minimum number of classes that would yield acceptable fit using χ^2 or the likelihood-ratio test. Alternatively, the choice of K can also be based on the information criteria such as the Akaiki information criteria (AIC; Akaike, 1987) and the Bayesian information criteria (BIC; Schwarz, 1978). In this approach, a number of models with different dimensions have to be estimated. Under some circumstances, it can be computationally ineffective and time consuming (Pan & Huang, 2014). Moreover, inference conditioning on a specific K from the 2-stage approach clearly ignores the uncertainty in the selection process (Yang et al., 2011).

Bayesian methods can potentially alleviates these difficulties. Unlike the above mentioned 2-stage approach, recent advances in Bayesian statistics allow joint inferences on the number of classes and model parameters. The development of the Markov chain Monte

Carlo (MCMC; refer to Neal, 1998 for a detailed review) methods provides powerful tools to perform posterior simulation. However, the traditional MCMC method is restricted to scenarios where the dimension of the parameter space is fixed. P. J. Green (1995) proposed a reversible jump Markov chain Monte Carlo (RJMCMC) method that offers a general framework for constructing reversible MCMC samplers that jump between between parameter spaces with different dimensions which enables the joint estimation of model dimensions and model parameters. Leveraging this technique, Pan and Huang (2014) developed a RJMCMC method for regression extension of the LCM. Similarly, Bartolucci, Farcomeni, and Scaccia (2017) considered a RJMCMC type sampler for the OLCM in the nonparametric IRT context. Alternative to the RJMCMC method, stochastic process priors can be used to specify a model with potentially an infinite number of dimensions. Among them, Dirichlet Process prior (Ferguson, 1973; Antoniak, 1974; Escobar & West, 1995; Neal, 2000; Navarro, Griffiths, Steyvers, & Lee, 2006) is a popular choice for partitioning data into an unknown number of clusters. For example, MacEachern (1994) deals with the problem of estimating a Gaussian mixture with unknown number of components, and Miyazaki and Hoshino (2009) develops a Bayesian nonparametric mixture IRT model. Different representations such as the Polya urn scheme and the stick-breaking process can be used to construct the Dirichlet process. However, the Chinese restaurant process (Blei, Griffiths, Jordan, & Tenenbaum, 2004; Griffiths & Ghahramani, 2011) is perhaps the most popular at this moment.

In this paper, we propose a modified Chinese restaurant process prior to accommodate the stochastic ordering constraints of the OLCM. In addition, we develop an efficient and easy to implement Gibbs sampler for posterior computation. Our method allows joint inference on the number of classes and item parameters for the OLCM.

4.2 Ordered latent class model with infinite classes

LCM assumes data arise from a mixture distribution. Its probability mass function can be described as

$$p(\boldsymbol{x}_i) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_i; \boldsymbol{\phi}_k), \tag{4.1}$$

where π_k is the marginal probability of allocation to the kth latent class, $p(\boldsymbol{x}_i; \boldsymbol{\phi}_k)$ gives the probability of $\boldsymbol{x}_i$ under the kth latent class, and K is the number of latent classes. In its marginal form as in (4.1), each observation is associated with parameters of all classes, i.e. $\boldsymbol{\phi}_k$ for $k = 1, 2, \ldots, K$. However, conditional on the latent class assignment $Z_i = k$, the probability of each observation depends only on the parameters specific to the kth latent class, i.e.

$$P(\boldsymbol{X}_i = \boldsymbol{x}_i | Z_i = k, \boldsymbol{\phi}) = P_{\boldsymbol{\phi}_k}(\boldsymbol{x}_i). \tag{4.2}$$

OLCM imposes ordering constraints on the parameters associated with different classes, specifically

$$\boldsymbol{\phi}_k \leq \boldsymbol{\phi}_{k'} \ \forall \ k < k'. \tag{4.3}$$

When the latent classes are ordered from low to high along the latent continuum, the probability of a correct response should be monotonically non-decreasing. It is similar to the monotonicity assumption in nonparametric IRT models (Sijtsma, 1998).

The generative process of the ordered latent class model can be represented using a Bayesian hierarchical model. Each subject belongs to one of the K latent classes whose distribution is governed by a categorical distribution with a probability vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$, i.e.

$$Z_i | \boldsymbol{\pi} \sim Categorical(\boldsymbol{\pi}). \tag{4.4}$$

Given $z_i = k$, the response vector from the ith subject follows a distribution parameterized by the parameters associated with the kth latent class, i.e.

$$\boldsymbol{X_i}|z_i, \boldsymbol{\phi} \sim F(\cdot|\boldsymbol{\phi}_k). \tag{4.5}$$

A natural choice for the prior of $\boldsymbol{\Pi}$ is the Dirichlet distribution, i.e.

$$\boldsymbol{\Pi} \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_K). \tag{4.6}$$

Respecting the stochastic ordering constraint in (4.3), the prior distribution of the parameters of the kth latent class should be truncated, i.e.

$$\boldsymbol{\Phi}_k \sim G_0(\cdot) I_{(l(\boldsymbol{\phi}_k), u(\boldsymbol{\phi}_k))}(\boldsymbol{\phi}_k), \tag{4.7}$$

where $I_{(l(\boldsymbol{\phi}_k), u(\boldsymbol{\phi}_k))}(\boldsymbol{\phi}_k)$ indicates the distribution is truncated to be in the region bounded by (l_j, u_j) for $j = 1, 2, \ldots, J$, and

$$l(\phi_{jk}) = \begin{cases} \phi_{jk-1} & k > 1 \\ -\infty & k = 1, \end{cases}$$

$$u(\phi_{jk}) = \begin{cases} \phi_{jk+1} & k < K \\ +\infty & k = K. \end{cases}$$

Equations (5) - (7) complete the specification of the Bayesian OLCM with K classes.

From (4.4), it is clear that the class assignments are conditionally independent given $\boldsymbol{\pi}$. It follows that integrating out $\boldsymbol{\pi}$ would induce dependencies between class assignments. Let $\boldsymbol{z}_{-i}$ denote the class assignments for all but the ith subject. The posterior distribution of the marginal class assignment probabilities conditional on the $N-1$ class assignments is

$$p(\boldsymbol{\pi}|\boldsymbol{z}_{-i}, \boldsymbol{\alpha}) \propto \prod_{i' \neq i} p(z_{i'}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}). \tag{4.8}$$

The standard categorical-Dirichlet conjugacy gives the closed form of (4.8),

$$\boldsymbol{\Pi}|\boldsymbol{z}_{-i}, \boldsymbol{\alpha} \sim Dirichlet(\alpha_1 + S_{-i,1}, \alpha_2 + S_{-i,2}, \ldots, \alpha_K + S_{-i,K}), \tag{4.9}$$

where $S_{-i,k}$ is the number of subjects (excluding the ith subject) classified into the kth class, i.e. $S_{-i,k} = \sum_{i' \neq i}[z_{i'} = k]$ and $[z_{i'} = k] = 1$ if $z_{i'} = k$, 0 otherwise. Integrating (4.4) over the above posterior distribution leads to

$$\begin{aligned} p(z_i|\boldsymbol{z}_{-i}, \boldsymbol{\alpha}) &= \int p(z_i|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{z}_{-i}, \boldsymbol{\alpha}) d\boldsymbol{\pi} && (4.10) \\ &= \int \pi_k \frac{1}{B(\boldsymbol{\alpha} + \boldsymbol{S}_{-i})} \prod_{j=1}^{K} \pi_j^{\alpha_j + S_{-i,j} - 1} d\boldsymbol{\pi} && (4.11) \\ &= \frac{B(\boldsymbol{\alpha} + \boldsymbol{S}_{-i} + \mathbf{1}^{(k)})}{B(\boldsymbol{\alpha} + \boldsymbol{S}_{-i})} && (4.12) \\ &= \frac{\alpha_k + S_{-i,k}}{\sum \alpha_j + N - 1}, && (4.13) \end{aligned}$$

where $\mathbf{1}^{(k)}$ is a $K-$length vector of zeros with a 1 in position k. From (12) to (13), it follows the definition of the Beta function, $B(\boldsymbol{y}) = \left(\prod \Gamma(y_i)\right) / \left(\Gamma\left(\sum y_i\right)\right)$, and the identity of the Gamma function, $\Gamma(y+1) = x\Gamma(y)$.

Often, there is no prior knowledge of favoring one class to another. In this case, a symmetric Dirichlet distribution can be used. That is $\alpha_1 = \alpha_2 = \cdots = \alpha_K = \alpha/K$. α is generally referred to as the concentration parameter since the distribution becomes more concentrated around the center of the $K-1$ simplex with a larger α and the distribution concentrates in the corners and along the boundaries of the simplex with a smaller α. The special case $\alpha/K = 1$ yields a uniform distribution. With the symmetric Dirichlet prior, (14) can be simplified as

$$p(z_i|\boldsymbol{z}_{-i}, \alpha, K) = \frac{\alpha/K + S_{-i,k}}{\alpha + N - 1}. \tag{4.14}$$

Equation (4.14) provides a finite dimensional prior that partitions data into K groups. It can be naturally extended to the infinite dimensional case by taking the limit $K \to \infty$ (Neal, 2000; Navarro et al., 2006). Consider the problem of assigning the ith subject into one of the infinitely many classes. The subject could fall into one of the classes that already has at least one member. The probability is given by

$$\begin{aligned} p(z_i = k|\boldsymbol{z}_{-i}, \alpha) &= \lim_{K\to\infty} \frac{\alpha/K + S_{-i,k}}{\alpha + N - 1} \\ &= \frac{S_{-i,k}}{\alpha + N - 1}, \end{aligned} \tag{4.15}$$

for k such that $S_{-i,k} > 0$. The subject could also fall into one of the infinitely many classes that yet has a member. Its probability is given by

$$\begin{aligned} p(z_i \in \{k; S_{-i,k} = 0\}|\boldsymbol{z}_{-i}, \alpha) &= 1 - \sum_{\forall k \ s.t. \ S_{-i,k}>0} \frac{S_{-i,k}}{\alpha + N - 1} \\ &= 1 - \frac{N-1}{\alpha + N - 1} \\ &= \frac{\alpha}{\alpha + N - 1}. \end{aligned} \tag{4.16}$$

Equations (4.15) and (4.16) defines a stochastic process prior that partitions the data into potentially infinitely many classes. Aldous (1985) provides an intuitive representation of the process - the Chinese restaurant process (CRP; see Griffiths & Ghahramani, 2011 for a detailed review). The CRP draws analogy from seating arrangements for customers of a Chinese restaurant. A customer can choose to sit at an existing table with $S_{-i,k}$ occupants. Alternatively, the customer can decide to sit at a new table without any prior occupants. The probability of sitting at an existing table is given by (4.15); while the probability of sitting at a new table follows (4.16). It should be noted that the prior probability of sitting at a new table (thus creating a new class) depends on the concentration parameter α. A

larger α encourages more classes. Coupled with an appropriate likelihood, the CRP can be used to define a wide range of Bayesian nonparametric models, i.e.

$$X_i|z_i = k, \phi_k \sim F(\cdot|\boldsymbol{\phi}_k)$$

$$\boldsymbol{Z}|\alpha \sim CRP(\cdot|\alpha)$$

$$\Phi_k \sim G_0(\cdot),$$

where F is the data model and G_0 is the prior for parameters associated with each class.

The CRP is closely related to the Dirichlet process and its other representations. Navarro et al. (2006) explains the CRP, the Polya urn scheme (Blackwell & MacQueen, 1973), the stick-breaking process (Sethuraman, 1994), and their connections to the Dirichlet process.

An important feature of the CRP prior is that the choice of the value for labeling different classes is completely arbitrary. Different sets of labels would lead to the same partition of the subjects as long as they consistently and faithfully describe whether subjects belong to the same class (Neal, 2000). From equation (4.16), it is clear that we don't need to distinguish which new table the customer chooses to sit. This will suffice in many applications of mixture models such as the LCM without the stochastic ordering constraint. However, it becomes a limitation in the specification of an OLCM. In this case, the ordering of the value of labels does not only signifies a partition of subjects, it also implies the stochastic ordering of the latent classes as in (4.3). As a result, the CRP as in (4.15) and (4.16) is not enough for the purpose of defining a prior for Bayesian nonparametric OLCM. Next, we introduce a modified CRP that can accommodate the stochastic ordering constraints of the latent classes.

To complete the motivation of our proposed prior, it is helpful to consider the following.

Let $\boldsymbol{\phi}'_i$ denote the parameter value associated with the ith observation. This is not to be confused with $\boldsymbol{\phi}_k$ which is the parameter assigned to all members of the kth class. Under the CRP, the prior induced for $\boldsymbol{\phi}'_i$ given the class assignments $z_i = k$, $\boldsymbol{z}_{-i}$ and the parameters associated with other observations $\boldsymbol{\phi}'_{-i}$ is

$$\boldsymbol{\Phi}'_i | z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\phi}'_{-i} \sim \begin{cases} G_0(\cdot) & S_{-i,k} = 0, \\ \delta(\cdot|\boldsymbol{\phi}_k) & S_{-i,k} \neq 0, \end{cases} \tag{4.17}$$

where $\delta(\cdot|\boldsymbol{\phi}_k)$ is a discrete distribution with a point mass at $\boldsymbol{\phi}'_i = \boldsymbol{\phi}_k$ and 0 everywhere else. Incorporating the stochastic ordering constraint, we have

$$\boldsymbol{\Phi}'_i | z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\phi}'_{-i} \sim \begin{cases} G_0(\cdot) I_{(l(\boldsymbol{\phi}_k), u(\boldsymbol{\phi}_k))}(\boldsymbol{\phi}_k) & S_{-i,k} = 0 \\ \delta(\cdot|\boldsymbol{\phi}_k) & S_{-i,k} \neq 0. \end{cases} \tag{4.18}$$

The truncation points are defined as

$$l(\boldsymbol{\phi}_k) = \begin{cases} \max\{\boldsymbol{\phi}'_j : z_j < k\} & \{\boldsymbol{\phi}'_j : z_j < k\} \neq \varnothing \\ -\infty & \{\boldsymbol{\phi}'_j : z_j < k\} = \varnothing, \end{cases} \tag{4.19}$$

and

$$u(\boldsymbol{\phi}_k) = \begin{cases} \min\{\boldsymbol{\phi}'_j : z_j > k\} & \{\boldsymbol{\phi}'_j : z_j > k\} \neq \varnothing \\ +\infty & \{\boldsymbol{\phi}'_j : z_j > k\} = \varnothing. \end{cases} \tag{4.20}$$

From (4.17) and (4.18), we can see the impact of enforcing the stochastic ordering constraint to the induced prior for the parameters associated with individual observations. If the observation gets assigned to a previously empty class, under the CRP, the partition of the subjects is invariant to the choice of the value of k; however, different choices of k could potentially induce different priors for $\boldsymbol{\phi}'_i$ when the stochastic ordering constraint is imposed.

This subtle but important difference suggests that the placement of the new tables needs to be distinguished to accommodate the stochastic ordering constraint. Furthermore, consider two previously empty tables, $k \neq k'$. If $l(\boldsymbol{\phi}_k) = l(\boldsymbol{\phi}_{k'})$ and $u(\boldsymbol{\phi}_k) = u(\boldsymbol{\phi}_{k'})$, the induced prior would be the same if the observation gets assigned to either k or k'. The above equalities will hold as long as the cardinalities of $\{j : k < z_j < k'\}$ and $\{j : k > z_j > k'\}$ are both zeros. In other words, when a customer chooses to sit at a new table, tables between two consecutive previously occupied tables are equivalent. It follows that, we should split the probability of sitting at a new table in a way so that it differentiates non-equivalent placements of the new table.

Observing the above properties, now we are ready to define the modified CRP formally. Let K^* denote the ordered set of currently occupied tables, i.e. $K^* = \{k_1^*, k_2^*, \ldots, k_{n^*}^*\}$ such that $k_u^* < k_t^*$ if $u < t$ and $S_{-i,k_u^*} \neq 0$ for $u = 1, 2, \ldots, n^*$. When the ith customer comes in, the probability of choosing a previously occupied table is

$$p(z_i = k_u^* | \boldsymbol{z}_{-i}, \alpha) = \frac{S_{-i,k_u^*}}{\alpha + N - 1}. \tag{4.21}$$

The probability of choosing a new table between two consecutive occupied tables is given by

$$p(z_i \in \{k; k_u^* < k < k_{u+1}^*\} | \boldsymbol{z}_{-i}, \alpha) = \frac{\alpha/(n^* + 1)}{\alpha + N - 1}, \tag{4.22}$$

for $u = 1, 2, \ldots, n^* - 1$. Notice that there are $n^* - 1$ such possibilities. A new table may also be chosen before the first occupied table with a probability of

$$p(z_i \in \{k; k < k_1^*\} | \boldsymbol{z}_{-i}, \alpha) = \frac{\alpha/(n^* + 1)}{\alpha + N - 1}. \tag{4.23}$$

Similarly, the probability of choosing a new table after the last occupied table is

$$p(z_i \in \{k; k > k_{n^*}^*\} | \boldsymbol{z}_{-i}, \alpha) = \frac{\alpha/(n^* + 1)}{\alpha + N - 1}. \tag{4.24}$$

Equations (4.21) to (4.24) complete the specification of the modified CRP prior. It is straightforward to verify that the marginal probability of choosing a new table is the same as (4.16). However, the modified CRP distinguishes among non-equivalent placements of the new table under the stochastic ordering constraint.

Using the modified CRP (MCRP), a Bayesian nonparametric OLCM can be specified as

$$X_i|z_i = k, \phi_k \sim F(\cdot|\phi_k), \tag{4.25}$$

$$\boldsymbol{Z}|\alpha \sim MCRP(\cdot|\alpha), \tag{4.26}$$

$$\boldsymbol{\Phi}_k \sim G_0(\cdot) I_{(l(\boldsymbol{\phi}_k), u(\boldsymbol{\phi}_k))}(\boldsymbol{\phi}_k). \tag{4.27}$$

4.2.1 Posterior computation

In a Bayesian context, we are interested in the posterior distribution $p(\boldsymbol{\phi}, \boldsymbol{z}|\boldsymbol{x})$. Exact computation of the posterior distribution for a Bayesian nonparametric OLCM is generally infeasible. Instead, we can sample from the posterior distribution using Markov chain Monte Carlo (MCMC) methods. Neal (2000) reviewed and developed a series MCMC algorithms for Dirichlet process mixture models. Building on the earlier work, we introduce a Gibbs sampler for Bayesian nonparametric OLCMs in this section.

In our Gibbs sampler, we repeatedly sample each z_i and $\boldsymbol{\phi}_k$ from their full conditional distributions. The full conditional distribution of class assignment for the ith individual is

$$p(z_i|\boldsymbol{z}_{-i}, \boldsymbol{x}_i, \boldsymbol{\phi}, \alpha) \propto p(z_i|\boldsymbol{z}_{-i}, \alpha) p(\boldsymbol{x}_i|z_i, \boldsymbol{\phi}), \tag{4.28}$$

where $P(z_i|\boldsymbol{z}_{-i}, \alpha)$ is given by the MCRP as in (4.21) through (4.24). If the ith individual gets assigned to a class already associated with other observations, the likelihood $P(\boldsymbol{x}_i|z_i, \boldsymbol{\phi})$ can

be evaluated under the item parameters from that class. However, if a new class is assigned, the evaluation would require integrating the item parameters over their priors. Specifically,

$$p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}_i, \boldsymbol{\phi}, \alpha) \propto \begin{cases} \frac{S_{-i,k}}{\alpha+N-1} f(\boldsymbol{x}_i|\boldsymbol{\phi}_k) & k \in K^* \\ \frac{\alpha/(n^*+1)}{\alpha+N-1} \int f(\boldsymbol{x}_i|\boldsymbol{\phi}_k) dG_0(\boldsymbol{\phi}_k) I_{(l(\phi_k),u(\phi_k))}(\phi_k) & k \in K'. \end{cases} \tag{4.29}$$

Even though it is useful in defining the MCRP, we cannot, of course, explicitly represent the infinite classes. Instead, we consider those classes currently associated with some observations, $k \in K^*$, and the $n^* + 1$ possible nonequivalent placements of the new class. In the above notation, $K' = \{k'_1, k'_2, \ldots, k'_{n^*+1}\}$ is an ordered set of such possible placements, and $k'_1 < k^*_1, k^*_1 < k'_2 < k^*_2, \ldots, k^*_n < k'_{n+1}$. The integral in (4.29) can be evaluated as

$$\int f(\boldsymbol{x}_i|\boldsymbol{\phi}_k) dG_0(\boldsymbol{\phi}_k) I_{(l(\phi_k),u(\phi_k))}(\phi_k) = \left(\frac{f(\boldsymbol{x}_i|\boldsymbol{\phi}_k) g_0(\boldsymbol{\phi}_k)}{G_0(u(\boldsymbol{\phi}_k)) - G_0(l(\boldsymbol{\phi}_k))} \right) / \left(\frac{h(\boldsymbol{\phi}_k|\boldsymbol{x}_i)}{H(u(\boldsymbol{\phi}_k)|\boldsymbol{x}_i) - H(l(\boldsymbol{\phi}_k)|\boldsymbol{x}_i)} \right), \tag{4.30}$$

where h is the posterior probability density function of $\boldsymbol{\phi}_k$ given the observation $\boldsymbol{x}_i$ before truncation, i.e.

$$h(\boldsymbol{\phi}_k|\boldsymbol{x}_i) = \frac{f(\boldsymbol{x}_i|\boldsymbol{\phi}_k) g_0(\boldsymbol{\phi}_k)}{\int f(\boldsymbol{x}_i|\boldsymbol{\phi}_k) g_0(\boldsymbol{\phi}_k) d\boldsymbol{\phi}_k}, \tag{4.31}$$

and H is the associated cumulative density function. When the ith individual is assigned to a previously empty class, i.e. $z_i = k, k \in K'$, the associated item parameters $\boldsymbol{\phi}_k$ should be drawn from the posterior distribution H.

Updating the item parameters $\boldsymbol{\phi}_k$ is relatively straightforward. The density of the full conditional distribution of $\boldsymbol{\phi}_k$ is given by

$$p(\boldsymbol{\phi}_k|\boldsymbol{z}_{-i}, \boldsymbol{x}) = b \prod_{i \in \{i; z_i = k\}} f(\boldsymbol{x}_i|\boldsymbol{\phi}_k) g_0(\boldsymbol{\phi}_k), \tag{4.32}$$

where b is an appropriate normalizing constant. The Metropolis algorithm (Metropolis et al., 1953, see Neal, 1998 for review) can be used if direct sampling from the above distribution is difficult.

We can summarize our Gibbs sampling method as follows:

Algorithm. Let the current state of the Markov chain consists of $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$ and $\boldsymbol{\phi} = \{\boldsymbol{\phi}_k; k \in K^*\}$. Repeatedly sample:

- For $i = 1, 2, \ldots, N$: If the class z_i has no other observation associated, i.e. $S_{-i,z_i} = 0$, remove $\boldsymbol{\phi}_{z_i}$ from $\boldsymbol{\phi}$ and z_i from K^*. Sample a new z_i from its full conditional distribution as in Equation (4.29). If the sampled new z_i is not associated with any other observation, sample $\boldsymbol{\phi}_{z_i}$ from $H(\cdot|\boldsymbol{x}_i)$ and add it to the state.

- For $k \in K^*$: Sample $\boldsymbol{\phi}_k$ from its full conditional distribution given all observations associated with class k as in Equation 4.29.

This algorithm is feasible as long as we can compute $\int f(\boldsymbol{x}_i|\boldsymbol{\phi}_k) dG_0(\boldsymbol{\phi}_k) I_{(l(\phi_k),u(\phi_k))}(\phi_k)$ and sample from H efficiently. This is generally the case when G_0 is conjugate to F, in which the closed form of H is available. For Dirichlet process mixture models with non-conjugate priors, Neal (2000) showed a Gibbs sampling method by augmenting auxiliary parameters. A similar strategy may be extended to the Bayesian nonparametric OLCMs.

4.2.2 The concentration parameter α

With a given sample size N, the choice of the concentration parameter α affects the prior probability of an observation being assigned to a previously empty class. In other words, the parameter indicates the level of ease of transition between dimensions (Miyazaki &

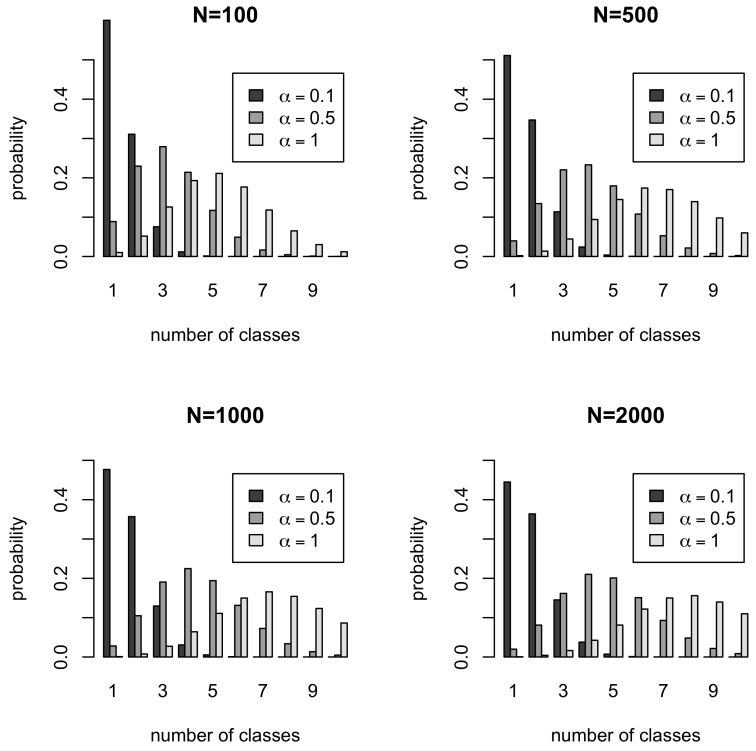


Figure 41: Prior probabilities $P(m|\alpha, N)$

Hoshino, 2009). A larger α encourages more classes. Selecting an appropriate value for the concentration parameter α is not always straightforward.

The probabilities of number of classes induced by the MCRP can be computed based on the results of Antoniak (1974). Let M denotes the cardinality of the set K^*, i.e. $M = |K^*|$. With N samples and the concentration parameter α, the probability of $M = m$ classes is given by

$$P(M = m|\alpha, N) = s(N, m)\alpha^m \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)}, \tag{4.33}$$

where $s(N, m)$ is the absolute value of the Stirling number of the first kind for N and m (Adamchik, 1997). Figure 41 shows the prior probability distribution of the number of classes with $\alpha = 0.1, 0.5, 1.0$ and $N = 100, 500, 1000, 2000$. The probabilities are computed

for $M = 1, 2, \ldots, 10$. For a small $\alpha = 0.1$, the prior distribution concentrates most of its probability mass on small m and decays quickly as the number of classes increases. In fact, almost 90% of its mass is on $m = 1$ and $m = 2$ for $N = 100$. A larger α shifts the mode of the probability distribution towards a larger number of classes. At the same time, the distribution tends to spread out more. For a fixed α, the MCRP prior allows more classes as the sample size increases.

In psychometrics, the appropriate number of latent classes is often reasonably moderate. For this reason, previous development of Dirichlet Process mixture IRT model employs a finite dimensional approach, and uses 10 as the upper limit of the number of components (Miyazaki & Hoshino, 2009). In our approach, $\alpha = 0.5$ seems to be a reasonable choice since it distributes adequate probability mass across $M = 1, 2, \ldots, 10$ under larger sample sizes.

A more flexible approach is to treat α as a random variable which can be learned from data. The conditional posterior of the concentration parameter α is

$$p(\alpha|m, N) \propto p(m|\alpha, N)p(\alpha), \tag{4.34}$$

where $p(m|\alpha, N)$ is given by the Antoniak equation as in Equation 4.33. Depending on the choice of the prior $p(\alpha)$, the exact form of the posterior distribution $p(\alpha|m, N)$ is generally unknown. However, it often can be approximated. If we are content with a limited range of α, a discretized version of $p(\alpha)$ can be used. As a result, the unnormalized posterior probabilities for each point in the range of α can be computed. Furthermore, the probabilities can be normalized easily. This is often referred to as the "griddy Gibbs" approach (Ritter & Tanner, 1992).

Alternatively, Escobar and West (1995) derived the exact posterior distribution for the

concentration parameter α when the prior $p(\alpha)$ comes from a gamma distribution. For $\alpha > 0$, the ratio of the gamma functions in Equation 4.33 can be written as

$$\frac{\alpha}{\alpha + N} = \frac{(\alpha + N)B(\alpha + 1, N)}{\alpha\Gamma(N)}, \tag{4.35}$$

where B is the beta function. Then the posterior distribution of α in Equation 4.34 becomes

$$p(\alpha|m, N) \propto p(\alpha)\alpha^{m-1}(\alpha + N)B(\alpha + 1, N), \tag{4.36}$$

$$\propto p(\alpha)\alpha^{m-1}(\alpha + N)\int_0^1 \eta^{\alpha}(1 - \eta)^{N-1}d\eta. \tag{4.37}$$

From equation 4.36 to equation 4.37, it follows the definition of the beta function. The above derivation implies that $p(\alpha|m, N)$ can be viewed as the marginal distribution from a joint distribution of α and a continuous random variable η whose support is between 0 and 1. Assuming a gamma prior, $p(\alpha) \propto \alpha^{a-1}e^{-b\alpha}$, the joint distribution is

$$p(\alpha, \eta|m, N) \propto \alpha^{a-1}e^{-b\alpha}\alpha^{m-1}(\alpha + N)\eta^{\alpha}(1 - \eta)^{N-1}. \tag{4.38}$$

However, direct sampling from the above joint distribution is difficult. Instead, we can sample α and η iteratively from their full conditional distributions. The full conditional for α is

$$p(\alpha|\eta, m, N) \propto \alpha^{a+m-2}e^{-b\alpha}(\alpha + N)\eta^{\alpha}, \tag{4.39}$$

$$\propto \alpha^{a+m-1}e^{-(b-\log\eta)\alpha} + N\alpha^{a+m-2}e^{-(b-\log\eta)\alpha}. \tag{4.40}$$

Multiplying a constant (w.r.t. α),$C\frac{(b-\log\eta)^{a+m}}{\Gamma(a+m)}\frac{(b-\log\eta)^{a+m-1}}{\Gamma(a+m-1)}$, to both summands, the density can be normalized as a mixture of two gamma densities, i.e.

$$\alpha|\eta, m, N \sim \pi Ga(a + m, b - \log\eta) + (1 - \pi)Ga(a + m - 1, b - \log\eta), \tag{4.41}$$

where $\pi/(1-\pi) = (a+m-1)/[N(b-\log\eta)]$. On the other hand, the full conditional distribution for η is

$$p(\eta|\alpha, m, N) \propto \eta^{\alpha}(1-\eta)^{N-1}, \tag{4.42}$$

which is $Beta(\alpha+1, N)$. Based on this result, α can be updated at the end of each Gibbs iteration by first drawing η from the beta distribution (4.42), then drawing α from the mixture gamma distribution (4.41).

4.2.3 Numerical demonstration

The utility of the proposed method in constructing ordered latent class models with unknown number of classes is demonstrated with a simple numerical example. We assign 1000 persons into two classes with class sizes being 0.30 and 0.70. In addition, we use 20 dichotomous items with item parameters 0.20 and 0.80 for the two classes. A 1000 by 20 dataset of binary responses are simulated. The Bayesian nonparametric OLCM is fitted for four different concentration parameters α - 0.001, 0.01, 0.1, and 0.5. The Markov chain is initialized with 3 classes having item parameters 0.5, 0.6, and 0.7. The 1000 persons are initially randomly assigned to the 3 classes with equal probabilities.

To summarize the results, we plot the posterior class sizes over 3000 posterior draws after the burn-ins in Figure 42. For readability, one sample is selected for every 60 draws. No matter the choice of α, the ordering of persons seems consistent. There is a clear separation of persons at the size 0.3. With a smaller α, the two classes are correctly identified. The sampler stays at the same dimension and rarely jumps to other dimensions. A lager α encourages more classes, it can be seen from the figure that more spurious classes exist

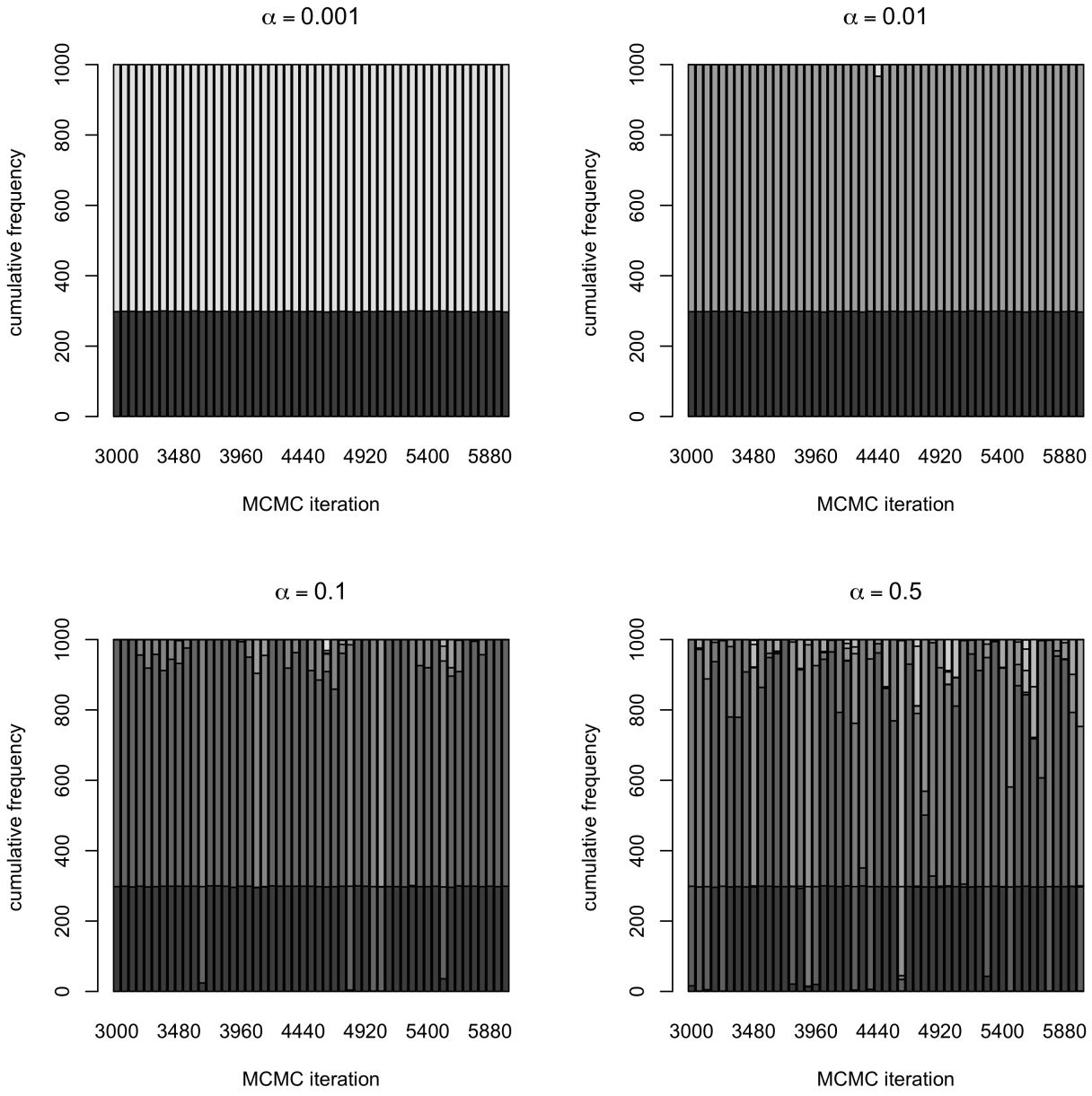


Figure 42: Posterior class sizes under fixed α

during each iteration. However, these classes are not persistent. They are born and die over the draws which means that these classes are not consistently occupied. In many applications of OLCM, in addition to scaling persons, interpreting the classes is often of interest. For this purpose, it is better to choose a smaller concentration parameter. It is less likely to have

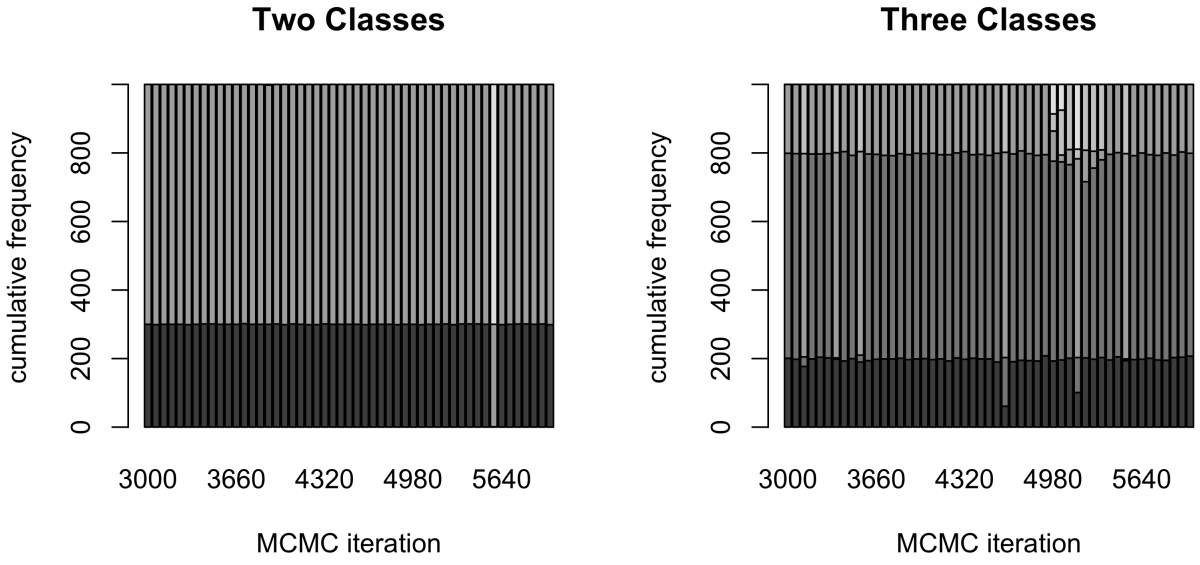


Figure 43: Posterior class sizes with $\alpha \sim Ga(0.001, 100)$

spurious classes in the posterior draws and much easier to summarize. In order to allow α to increase along the sample size N, some authors suggest $1/N$ as a default choice (Escobar & West, 1995).

Besides fixing α, we can also put a prior on the concentration parameter. For this demonstration, I chose $Ga(0.001, 100)$. Figure 43 shows the posterior class sizes. In addition to the two class condition, I also simulated a condition where the true number of classes is three with class sizes being 0.20, 0.60, and 0.20. The item parameters are 0.1, 0.5, and 0.9 for all 20 items. Figure 44 shows the trace plot of the posterior samples of α. It is interesting that, under the same Gamma prior, the posterior samples of α tends to be slightly larger for the 3 classes data. It suggests that the data does contain information about the concentration parameter α.

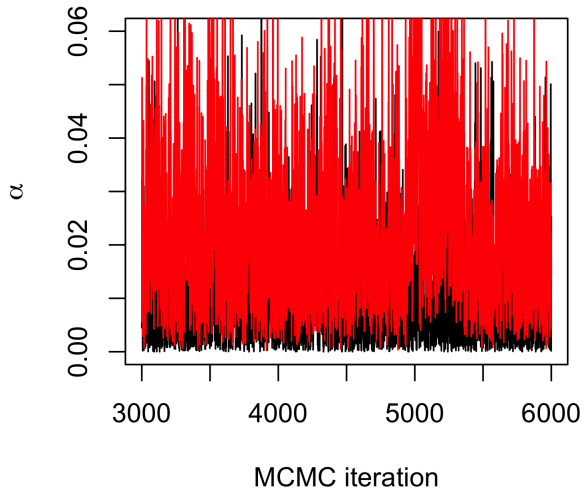


Figure 44: Trace lines for posterior samples of α - black: 3 classes, red: 4 classes

4.3 Simulations

In this section, we demonstrate the application of the proposed method in the context of estimating the item response function (IRF). Let X_{ij} denote the response from the ith individual to the jth item for $i \in \{1, 2, \ldots, N\}$ and $j \in \{1, 2, \ldots, J\}$. Under the unidimensional IRT, the item response random variables $X_{i1}, X_{i2}, \ldots, X_{iJ}$ are assumed to be conditionally independent given the latent variable θ_i. Furthermore, the IRF specifying the response probabilities given the latent variable, i.e. $P_j(\theta_i) = P(X_{ij} = 1|\theta_i)$ in the dichotomous case, is constrained to be non-decreasing. Formally, for any $\theta_i > \theta_{i'}$, $P(X_{ij} = 1|\theta_i) \geq P(X_{i'j}|\theta_{i'})$. In addition, the latent variable θ is often assumed to follow some distribution $F(\theta)$. A common choice is the standard normal distribution, $F(\cdot) = \Phi(\cdot)$.

Parametric IRT models, such as the Rasch model, the two-parameter logistic (2PL) model, and the three-parameter logistic (3PL) model, give specific parametric form to the

IRF. These class of models are relatively easy to fit at the expense of being less flexible than nonparametric approaches. The Bayesian nonparametric OLCM introduced in this paper provides a flexible yet computationally straightforward approach to estimate IRFs. We first simulate the case where the distribution $F(\cdot)$ is discrete.

4.3.1 Study 1: discrete $F(\cdot)$

Templin and Bradshaw (2014) proposed a hierarchical diagnostic classification model (HDCM) and analyzed the Examination for the Certificate of Proficiency in English (ECPE) dataset. There are 28 items measuring 3 attributes - morphosyntactic (α_1), cohesive (α_2), and lexical (α_3). The HDCM hypothesizes a linear hierarchy among the three attributes, i.e. $\alpha_3 \rightarrow \alpha_2 \rightarrow \alpha_1$. Following this attribute hierarchy, an individual cannot have mastered α_1 without having mastered α_2. Similarly, the attribute α_2 is a prerequisite for α_3. As a result, the number of admissible latent class reduces from $2^3 = 8$ to 4. More importantly, because of the hierarchical relationship, the four latent classes are fully ordered. For example, an individual with an attribute pattern $(0, 0, 1)$ cannot have higher probability of getting a correct response on any item than another individual with the attribute pattern $(0, 1, 1)$. We should also note a subtle difference between the OCLM and the HDCM. The Q-matrix in the HDCM specifies equivalences between some latent classes on each item. For instance, the first item in the dataset measures α_1 and α_2. In other words, the state of mastery of α_3 has no impact on the response probability for that item. As a result, the response probabilities between the class $(0, 0, 0)$ and $(0, 0, 1)$ are constrained to be the same.

Based on the item parameters of the HDCM reported in Templin and Bradshaw (2014),

we computed item response probabilities for each of the four classes (see Table 41). Templin

Table 41: ECPE item response probabilities under HDCM

	Class 1 $(0,0,0)$	Class 2 $(0,0,1)$	Class 3 $(0,1,1)$	Class 4 $(1,1,1)$
item 1	0.705	0.705	0.815	0.926
item 2	0.746	0.746	0.904	0.904
item 3	0.421	0.507	0.507	0.791
item 4	0.460	0.822	0.822	0.822
item 5	0.743	0.955	0.955	0.955
item 6	0.701	0.925	0.925	0.925
item 7	0.490	0.707	0.707	0.944
item 8	0.822	0.822	0.964	0.964
item 9	0.527	0.786	0.786	0.786
item 10	0.527	0.527	0.527	0.898
item 11	0.495	0.719	0.719	0.918
item 12	0.143	0.389	0.389	0.745
item 13	0.668	0.668	0.668	0.913
item 14	0.552	0.552	0.552	0.835
item 15	0.723	0.956	0.956	0.956
item 16	0.488	0.692	0.692	0.911
item 17	0.801	0.863	0.943	0.943
item 18	0.711	0.909	0.909	0.909
item 19	0.445	0.836	0.836	0.836
item 20	0.199	0.392	0.392	0.767
item 21	0.547	0.788	0.788	0.918
item 22	0.289	0.792	0.792	0.792
item 23	0.670	0.670	0.935	0.935
item 24	0.350	0.350	0.694	0.694
item 25	0.527	0.527	0.527	0.784
item 26	0.537	0.781	0.781	0.781
item 27	0.306	0.306	0.306	0.705
item 28	0.632	0.908	0.908	0.908

and Bradshaw (2014) also reported the estimated class sizes to be 0.320, 0.144, 0.184, and 0.351. According to these class sizes, 1000 individuals are randomly allocated to one of the four classes. Using the computed item response probabilities, we generated 100 datasets of item responses from the $N = 1000$ individuals to the $J = 28$ items.

For each generated dataset, we fitted the Bayesian nonparametric OLCM with the con-

centration parameter $\alpha = 0.5$. To ensure convergence, 6000 MCMC iterations are used with the first 3000 treated as burn-ins. Within the tth iteration of the posterior sampling, the MCMC algorithm samples the class assignment $z_i^{(t)}$ for $i = 1, 2, \ldots, N$ and the item response probabilities for each non-empty class, $\boldsymbol{\phi}_k^{(t)}$ $\forall k \in K^{*(t)}$. By assuming a unidimensional latent trait following the standard normal distribution, $\theta \sim \Phi(\cdot)$, a continuous IRF can be fitted. The corresponding latent class assignment at the ability level θ can be found by computing $z_{\theta^*} = F^{-1}(\Phi(\theta))$, where $F(z) = P(Z \leq z) = \sum_{k \in K^*: k \leq z} f(k)$ is the cumulative distribution function for the class assignment. It follows that the IRF is $P_j(\theta) = \phi_{jz_\theta}$. In this simulation, we computed IRF for θ on a finite grid from -4.0 to 4.0 by increment of 0.1. Averaging the fitted IRF over the posterior samples leads to the expected a posteriori (EAP) estimate of the IRF, $\hat{P}_j(\theta) = (1/T) \sum_t P_j^{(t)}(\theta)$. For comparison, we also fitted the 2PL on the same generated datasets.

To evaluate the effectiveness in estimating the IRF, we propose a criterion (CR) that computes the expected absolute difference between the estimated item response probability and the true item response probability over the distribution $\Phi(\cdot)$, i.e.

$$CR = \int_\theta |\hat{P}_j(\theta) - P_j(\theta)| d\Phi(\theta). \tag{4.43}$$

In practice, CR is approximated using quadrature points of θ defined earlier, specifically,

$$CR = \sum_m \left(|\hat{P}_j(\theta_m) - P_j(\theta_m)| [\Phi(\theta_m + 0.05) - \Phi(\theta_m - 0.05)] \right). \tag{4.44}$$

Table 42 shows the mean and the variance of the average CR of the 28 items over the 100 datasets. Based on CR, Bayesian nonparametric OLCM estimates the IRFs consistently better than the 2PL. Figure 45 demonstrates the mean of the estimated IRFs over the 100 datasets for item 10 and item 20. From the illustration, we can see that Bayesian

Table 42: Summary of the average CR

	BNP OLCM	2PL
mean	3.9431×10^{-2}	$6.7771 \times 10{-}2$
variance	8.1583×10^{-6}	1.4274×10^{-7}

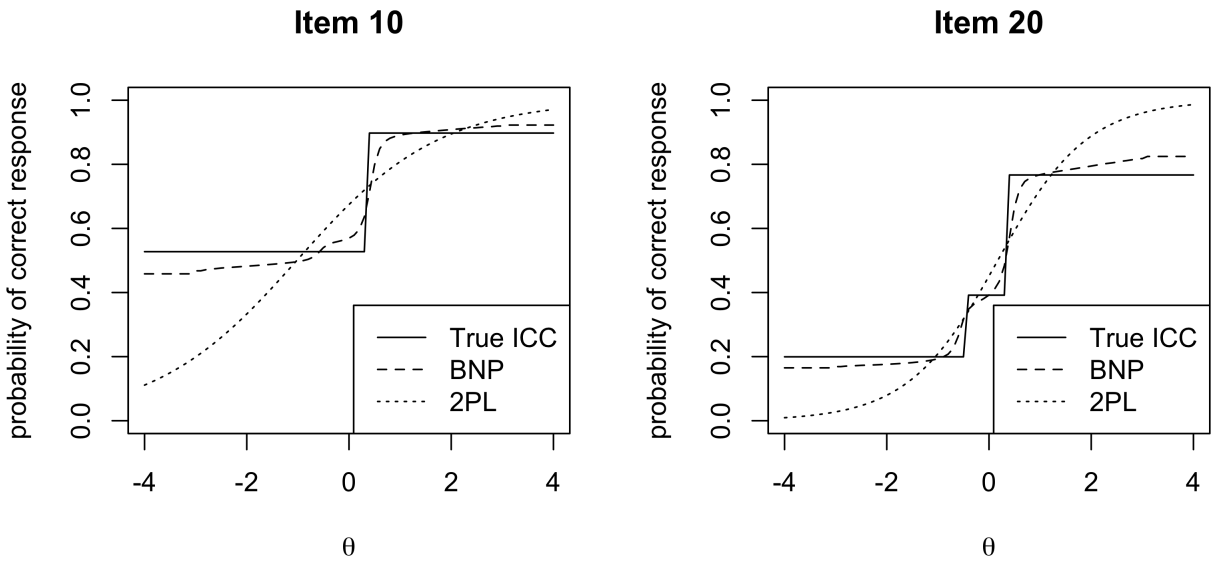


Figure 45: Mean estimated IRFs in simulation study 1

nonparametric OLCM is very flexible in estimating the IRFs across a wide range of θ; while the 2PL, on the other hand, is not as nearly flexible. This is especially obvious as θ moves towards either extreme where the 2PL grossly overestimates or underestimates the true response probabilities. Figure 46 shows the mean absolute difference between the estimated IRF and true IRF averaged over the 28 items. The Bayesian nonparametric OLCM approach uniformly outperforms the 2PL across the entire range of θ.

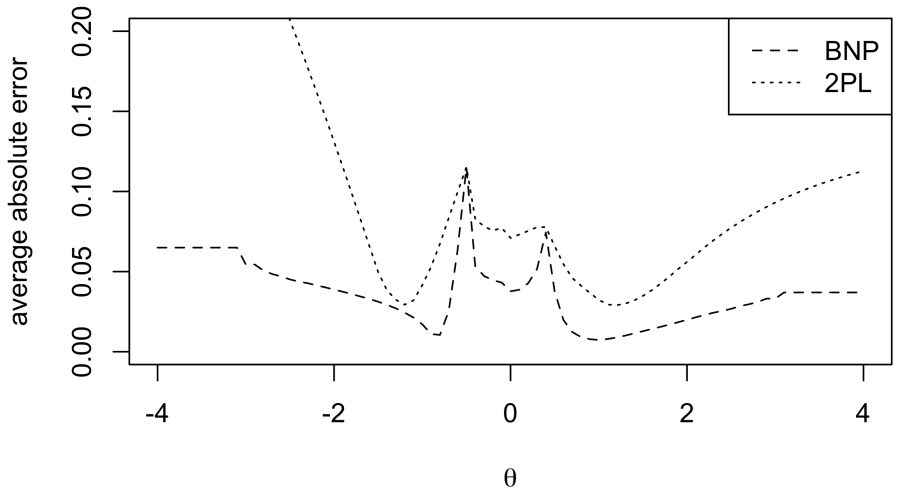


Figure 46: Mean of $|\hat{p}_j(\theta) - p_j(\theta)|$ over 28 items in simulation study 1

4.3.2 Study 2: continuous $F(\cdot)$

The mediocre performance of the 2PL in the the first simulation is anticipated since the parametric assumption of the model is violated. In this second simulation study, we investigate the case where the parametric assumption of the 2PL is met.

We randomly generated $J = 20$ pairs of item discrimination and difficulty parameters from their usual ranges, i.e. $a_j \sim unif(0.5, 2.0)$ and $b_j \sim unif(-3.0, 3.0)$ for $j = 1, 2, \ldots, J$ (see table 43). The latent abilities for $N = 1000$ individuals are generated from the standard normal distribution, i.e. $\theta_i \sim \Phi$ for $i = 1, 2, \ldots, N$. The item response probabilities are computed according to the 2PL,

$$P_j(\theta_i) = P(X_{ij} = 1|\theta_i) = \frac{\exp(a_j(\theta_i - b_j))}{1 + \exp(a_j(\theta_i - b_j))}. \tag{4.45}$$

Using these response probabilities, we generated 100 datasets of binary responses. Similar to the simulation study 1, for each dataset, we fitted the Bayesian nonparametric OLCM

Table 43: Generated item parameters in simulation study 2

	a	b
item 1	1.451	-0.766
item 2	1.665	-2.095
item 3	1.372	0.827
item 4	1.336	0.227
item 5	0.552	1.414
item 6	1.054	1.146
item 7	1.790	0.955
item 8	1.219	0.120
item 9	0.621	-1.864
item 10	1.905	0.100
item 11	1.146	2.432
item 12	1.143	-0.954
item 13	0.825	0.299
item 14	1.972	2.043
item 15	0.926	-0.056
item 16	1.586	-1.150
item 17	1.307	-0.300
item 18	1.566	1.164
item 19	1.813	1.444
item 20	1.502	-1.137

and the 2PL.

Table 44 shows the mean and the variance of CR for both the Bayesian nonparametric

Table 44: Summary of the average CR in simulation study 2

	BNP OLCM	2PL
mean	2.3563×10^{-2}	1.4466×10^{-2}
variance	2.0639×10^{-5}	9.3416×10^{-6}

OLCM and the 2PL. Not surprisingly, in this case, the 2PL performs better estimating the IRFs. But the Bayesian nonparametric OLCM approach does not fare much worse. In fact, considering both the first and the second simulation studies, our proposed method estimates IRFs consistently well regardless of the parametric forms of the true IRFs. Figure 47 plots

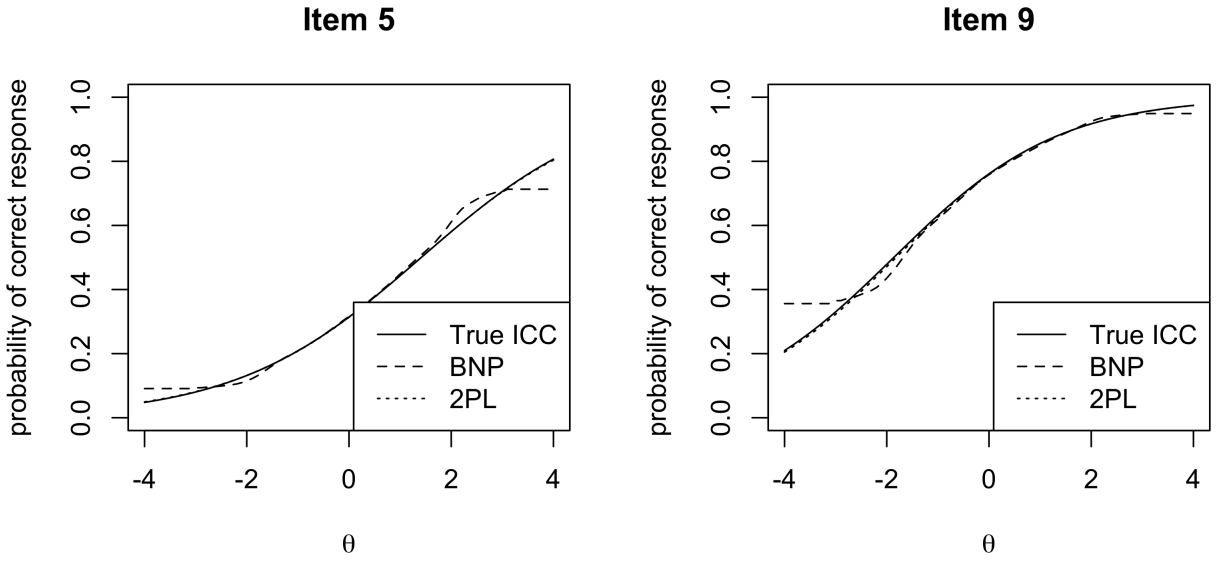


Figure 47: Mean estimated IRFs in simulation study 2

the average of the estimated IRFs for both approaches. Examining the estimated item characteristic curves (ICC), we found that both the 2PL and our approach estimate the IRFs very well for the middle range of θ. As θ goes to either extreme, the estimated IRFs produced by the Bayesian nonparametric OLCM deviate a little from the true IRFs. However, this is very much expected since there are very low probability density with θ greater than 3.0 or less than -3.0 within the standard normal distribution. For the Bayesian nonparametric approach, more accurate estimation of IRFs for those θ values requires a much larger sample size. Figure 48 further reveals that the two approaches are very comparable in estimating the IRFs for a good range of θ.

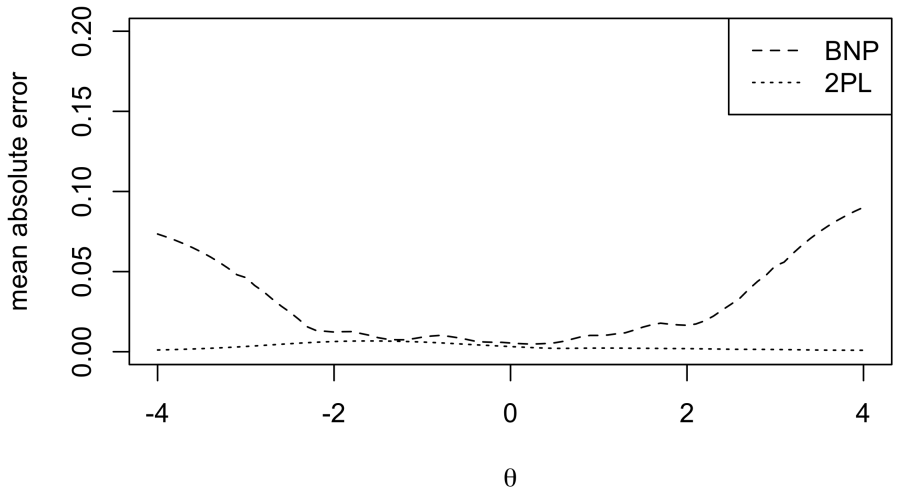


Figure 48: Mean of $|\hat{p}_j(\theta) - p_j(\theta)|$ over 20 items in simulation study 2

4.4 Real data analysis

A common application of nonparametric IRT is to assess the attainability of the parametric model assumption (i.e. Lee, Wollack, & Douglas, 2009). In this section, we demonstrate an application of the proposed Bayesian nonparametric OLCM in checking model fit by analyzing the ECPE dataset. The ECPE dataset is publicly available and can be accessed through the *R* "CDM" package (George et al., 2016). The data contains the binary responses of 2922 persons answering 28 items measuring three attributes as described earlier. We fit the Bayesian nonparametric OLCM and compare the results with those of Templin and Bradshaw (2014).

Figure 49 depicts average absolute differences between the estimated IRFs of the Bayesian OLCM and the HDCM. For the middle range of θ, the difference is generally under 0.10. However, as θ moves towards the negative extreme, the difference is getting large very quickly.

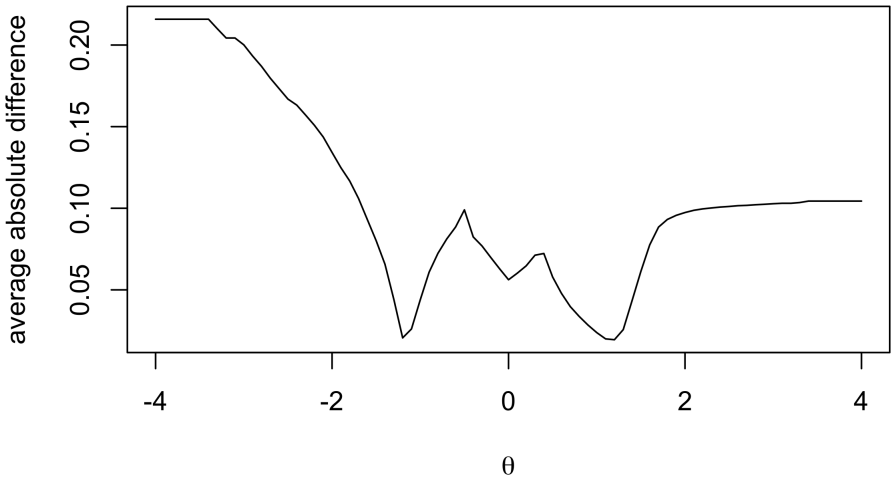


Figure 49: Average absolute difference of the estimated IRFs for the ECPE data

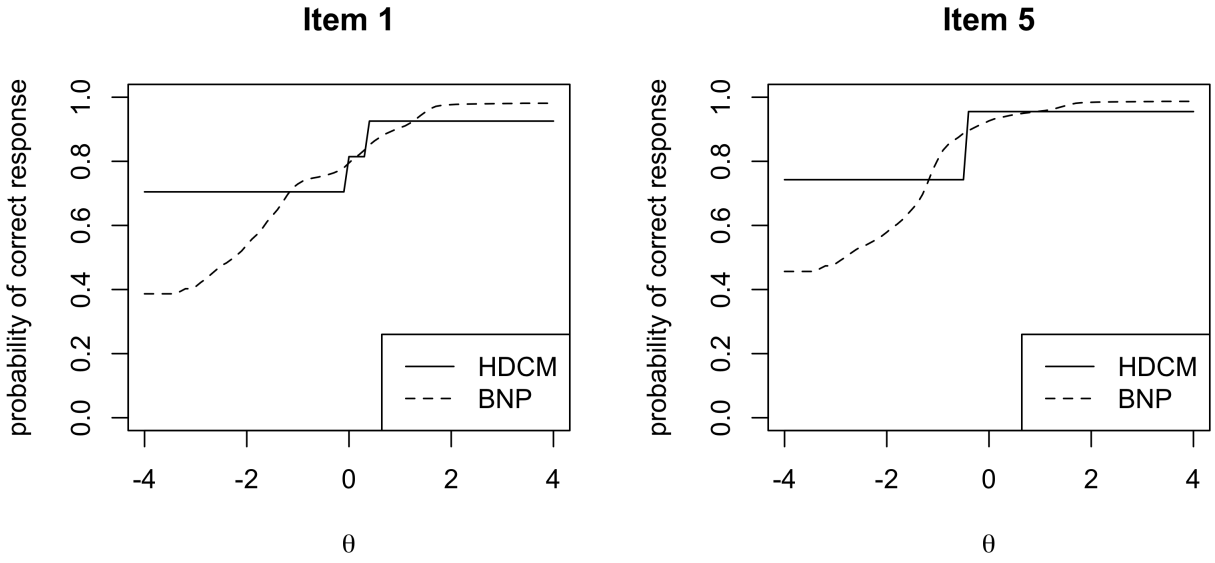


Figure 410: Estimated IRFs for the ECPE dataset

For an extreme negative θ, the difference could be an alarming 0.20.

We further examined the estimated IRFs for each item. Figure 410 provides two examples. While the HDCM estimates IRFs reasonably close to the Bayesian nonparametric

OLCM for the middle and upper level of θ, it overestimates the item responses consistently for the lower level of θ. Put this results into perspective, under the HDCM, a person with none of the required attributes has over 0.7 probability of getting an item correct. At the same time, the data seems to encourage further distinction of these persons. In other words, if we believe there is an attribute hierarchy, some lower level attribute(s) may be potentially missing from the Q-matrix.

4.5 Discussion

While parametric item response models such as the 2PL are generally more popular, the nonparametric method provides a flexible tool for analyzing a wide range of data. There are many different approaches in developing nonparametric item response models. For example, Johnson (2007) introduced a free-knot splines nonparametric model. The Bayesian nonparametric OLCM, on the other hand, takes a ordered latent class approach to nonparametric item response modeling. Compared to previous development of nonparametric IRT models, the current model is conceptually more straightforward and easier to implement. In addition, the proposed model and estimation algorithm can be readily adapted to response types beyond the dichotomous case.

The nonparametric models are certainly not without their drawbacks. For example, it is often difficult to scale items under the nonparametric models. Moreover, the estimation of the nonparametric IRT models are generally more complex and time-consuming. While the MCMC algorithm described in this paper can handle moderate sample sizes efficiently, development of fast posterior approximation algorithms could be very useful dealing with

more complex and larger data in the future.

Chapter 5

Thoughts on Future Research

The papers in this dissertation have touched some of the most popular latent variable models. The first paper deals with the classic inference problem of the interval estimation of the continuous latent trait under the IRT models. The latent variable could also be discrete. CDM, discussed in the second paper, is a type of restricted latent class models where the item parameters associated with different classes are restricted through the Q-matrix (Xu & Shang, 2018). The third paper further considers the ordered latent class models. Moreover, the infinite class ordered latent class model can be used to approximate a single continuous latent variable. Besides the three papers, I have also studied other topics during my graduate study. Some of them are closely related to this dissertation. One example is the development of the power divergence family of statistics for person parameters in IRT models.

Testing binomial proportions is a classic statistical inference problem. Many well-known statistics have been developed for this purpose. Some of them include the log-likelihood ratio statistic (G^2), Person's Chi-square statistic (χ^2), the Freeman-Tukey statistic (T^2), the Neyman modified Chi-square statistic (NM^2) and the modified log-likelihood ratio statistic

(MG^2). But, more importantly, they can be studied under the same family. Cressie and Read (1984) introduced the power divergence family of statistics in their seminal paper where the aforementioned statistics are special cases. This development provides a powerful framework for other researchers to study the statistical properties of binomial inference problems in depth (e.g. Jin, Thulin, & Larsson, 2017). Extending the power divergence family to IRT models could be a very significant contribution.

There is a subtle but important difference between a binomial model and an IRT model. Given the proportion (or probability of a success), each observation is independently and identically distributed under the binomial model. However, conditional on the person parameter, item responses are independently but, in general, not necessarily identically distributed under the IRT models. Thus, extending the power divergence family framework to IRT is not direct and trivial. In one of my most recent papers (X. Liu, Yang, Chae, & Natriello, 2018), we proposed an extension of the power divergence (PD) family of statistics to IRT models. Furthermore, we prove the asymptotic equivalence among all statistics in the PD family and they have the same χ_1 limiting distribution. we also extract higher order error terms of approximating the exact distribution of the statistics using the limiting distribution which enables us to compare statistics within the family.

One feature (or rather limitation) of these research is that they only model the primary outcome data - student responses to test items. Traditionally, large-scale assessments were administered in a pencil-and-paper format. As a result, very limited information can be recorded besides the primary item responses. Thanks to the advancement of technology, assessments become more and more technology-based. Software can be designed to log students activity within the system at a much finer grain size (for a review, see Bergner &

von Davier, 2018). The data collected are usually referred to as the process data. How the process data can be utilized to better understand or improve measurement is a challenge and opportunity to the field of psychometrics.

References

Adamchik, V. (1997). On Stirling numbers and Euler sums. *Journal of Computational and Applied Mathematics*, *79*(1), 119–130. Retrieved from https://www.sciencedirect.com/science/article/pii/S0377042796001677 doi: 10.1016/S0377-0427(96)00167-7

Agresti, A. (2003). Dealing with discreteness: making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical methods in medical research*, *12*(1), 3–21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12617505

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332. doi: 10.1007/BF02294359

Aldous, D. J. (1985). Exchangeability and related topics. In P. L. Hennequin (Ed.), *École d'été de probabilités de saint-flour xiii — 1983* (pp. 1–198). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/BFb0099421 doi: 10.1007/BFb0099421

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, *2*(6), 1152–1174. Retrieved from http://projecteuclid.org/euclid.aos/1176342871 doi: 10.1214/aos/1176342871

Baker, F. B., & Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques* (2nd ed.). Boca Raton, FL: CRC Press.

Bartolucci, F., Farcomeni, A., & Scaccia, L. (2017). *A Nonparametric Multidimensional Latent Class IRT Model in a Bayesian Framework.* Springer US. Retrieved from http://link.springer.com/10.1007/s11336-017-9576-7 doi: 10.1007/s11336-017-9576-7

Bergner, Y., & von Davier, A. A. (2018). Process Data in NAEP. *Journal of Educational and Behavioral Statistics*, 107699861878470. Retrieved from http://journals.sagepub.com/doi/10.3102/1076998618784700 doi: 10.3102/1076998618784700

Biehler, M., Holling, H., & Doebler, P. (2014). Saddlepoint Approximations of the Distribution of the Person Parameter in the Two Parameter Logistic Model. *Psychometrika*, *80*(3), 665–688. Retrieved from http://dx.doi.org/10.1007/s11336-014-9405-1 doi: 10.1007/s11336-014-9405-1

Blackwell, D., & MacQueen, J. B. (1973). Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, *1*(2), 353–355. Retrieved from http://projecteuclid.org/euclid.aos/1176342372 doi: 10.1214/aos/1176342372

Blei, D., Griffiths, T., Jordan, M., & Tenenbaum, J. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, *16*, 106. Retrieved from http://scholar.google.com/scholar?hl=en{\&}btnG=Search{\&}q=intitle:Hierarchical+Topic+Models+and+the+Nested+Chinese+Restaurant+Process{\#}0 doi: 10.1016/0169-023X(89)90004-9

Bock, D. R., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*(2), 179–197. Retrieved from http://link.springer.com/10.1007/BF02291262 doi: 10.1007/BF02291262

Brent, R. P. (1973). *Algorithms for Minimization without Derivatives*. Englewood Cliffs, New Jersey: Prentice-Hall.

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. doi: 10.1080/10618600.1998.10474787

Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (2011). Handbook of Markov Chain Monte Carlo. *Handbook of Markov Chain Monte Carlo*. doi: 10.1201/b10905

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*(1), 33–57. doi: 10.1007/s11336-009-9136-x

Casella, G., & Berger, R. (2001). *Statistical Inference*. Duxbury Resource Center. Textbook Binding.

Cha, J. (2011). *Application of ordered latent class regression model in educational assessment* (Doctoral dissertation, Columbia University). Retrieved from https://doi.org/10.7916/D8ZS33QF

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian Estimation of the DINA Q matrix. *Psychometrika*, *83*(1), 89–108. Retrieved from http://link.springer.com/10.1007/s11336-017-9579-4 doi: 10.1007/s11336-017-9579-4

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical Analysis of Q-Matrix

Based Diagnostic Classification Models. *Journal of the American Statistical Association*, *110*(510), 850–866. Retrieved from http://www.tandfonline.com/doi/full/10.1080/01621459.2014.934827 doi: 10.1080/01621459.2014.934827

Chung, M.-t. (2014). *Estimating the Q-matrix for Cognitive Diagnosis Models in a Bayesian Framework* (Doctoral dissertation, Columbia University). Retrieved from https://search.proquest.com/docview/1548332406

Cressie, N., & Read, T. R. C. (1984). Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46*(3), 440–464. Retrieved from http://www.jstor.org/stable/2345686 doi: 10.2307/2345686

Croon, M. (1990). Latent class analysis with ordered latent classe. *British Journal of Mathematical and Statistical Psychology*, *43*(2), 171–192. Retrieved from http://doi.wiley.com/10.1111/j.2044-8317.1990.tb00934.x doi: 10.1111/j.2044-8317.1990.tb00934.x

Culpepper, S. A. (2015). Bayesian Estimation of the DINA Model With Gibbs Sampling. *Journal of Educational and Behavioral Statistics*, *40*(5), 454–476. Retrieved from http://jeb.sagepub.com/cgi/doi/10.3102/1076998615595403 doi: 10.3102/1076998615595403

DeCarlo, L. T. (2012). Recognizing Uncertainty in the Q-Matrix via a Bayesian Extension of the DINA Model. *Applied Psychological Measurement*, *36*(6), 447–468. Retrieved from http://journals.sagepub.com/doi/10.1177/0146621612449069 doi: 10.1177/0146621612449069

DeCarlo, L. T., & Kinghorn, B. R. (2016). An Exploratory approach to the Q-matrix via Bayesian estimation. In *Paper presented at the 2016 meeting of the national council on measurement in educaiton, washington, dc.*

de la Torre, J. (2008). DINA Model and Parameter Estimation: A Didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130. Retrieved from http://jeb.sagepub.com/cgi/doi/10.3102/1076998607309474 doi: 10.3102/1076998607309474

de la Torre, J. (2011). The Generalized DINA Model Framework. *Psychometrika*, *76*(2), 179–199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. Retrieved from http://link.springer.com/10.1007/BF02295640 doi: 10.1007/BF02295640

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Se-*

ries B, *39*(1), 1–38. Retrieved from http://www.jstor.org/stable/2984875 doi: 10.2307/2984875

Doebler, A., Doebler, P., & Holling, H. (2012). Optimal and Most Exact Confidence Intervals for Person Parameters in Item Response Theory Models. *Psychometrika*, *78*(1), 98–115. Retrieved from http://dx.doi.org/10.1007/s11336-012-9290-4 doi: 10.1007/s11336-012-9290-4

Escobar, M. D., & West, M. (1995). Bayesian Density Estimation and Inference using Mixtures. *Journal of the american statistical*, *90*(430), 577–588. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476550

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, *1*(2), 209–230. Retrieved from http://projecteuclid.org/euclid.aos/1176342360 doi: 10.1214/aos/1176342360

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10476213 doi: 10.1080/01621459.1990.10476213

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*. doi: 10.1214/ss/1177011136

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22499653 doi: 10.1109/TPAMI.1984.4767596

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R Package CDM for Cognitive Diagnosis Models. *Journal of Statistical Software*, *74*(2). Retrieved from http://www.jstatsoft.org/v74/i02/ doi: 10.18637/jss.v074.i02

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4*. doi: 1176289

Goodman, L. a. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231. doi: 10.2307/2334349

Green, B. (1951). A general solution for the latent class model of latent structure analysis.

Psychometrika, *16*(2), 151–166. Retrieved from https://ideas.repec.org/a/spr/psycho/v16y1951i2p151-166.html doi: 10.1007/BF02289112

Green, P. J. (1995). Reversible jump Markov chain monte carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732. Retrieved from https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/82.4.711 doi: 10.1093/biomet/82.4.711

Griffiths, T. L., & Ghahramani, Z. (2011). The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, *12*, 1185–1224. Retrieved from https://cocosci.berkeley.edu/tom/papers/indianbuffet.pdf doi: 10.1016/j.biotechadv.2011.08.021.Secreted

Hagell, P., & Westergren, A. (2011). Measurement Properties of the SF-12 Health Survey in Parkinson's Disease. *Journal of Parkinson's Disease*, *1*, 185–196. Retrieved from http://content.iospress.com.ezproxy.cul.columbia.edu/download/journal-of-parkinsons-disease/jpd11026?id=journal-of-parkinsons-disease{\%}2Fjpd11026 doi: 10.3233/JPD-2011-11026

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing. doi: 10.1007/978-94-017-1988-9

Hartz, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory With Practicality* (Doctoral dissertation, University of Illinois at Urbana-Champaign). Retrieved from http://hdl.handle.net/2142/87393

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, *57*(1), 97–109. Retrieved from http://www.jstor.org/stable/2334940

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables. *Psychometrika*, *74*(2), 191–210. Retrieved from http://link.springer.com/10.1007/s11336-008-9089-5 doi: 10.1007/s11336-008-9089-5

Jin, S., Thulin, M., & Larsson, R. (2017). Approximate Bayesianity of Frequentist Confidence Intervals for a Binomial Proportion. *American Statistician*, *71*(2), 106–111. doi: 10.1080/00031305.2016.1208630

Johnson, M. S. (2004). Item response models and their use in measuring food insecurity and hunger. In *Paper presented at the workshop on the measurement of food insecurity and hunger. the national academy of science panel to review usda's measurement of food insecurity and hunger.*

Johnson, M. S. (2007). Modeling dichotomous item responses with free-knot splines.

Computational Statistics and Data Analysis, *51*(9), 4178–4192. Retrieved from www.elsevier.com/locate/csda doi: 10.1016/j.csda.2006.04.021

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, *25*(3), 258–272. Retrieved from http://journals.sagepub.com/doi/10.1177/01466210122032064 doi: 10.1177/01466210122032064

Klauer, K. C. (1991). Exact and best confidence intervals for the ability parameter of the Rasch model. *Psychometrika*, *56*(3), 535–547. Retrieved from http://dx.doi.org/10.1007/BF02294489 doi: 10.1007/BF02294489

Land, A. H., & Doig, A. G. (1960). An automatic method of solving discrete programming problems. *Econometrica*, *28*(3), 497–520. Retrieved from http://www.jstor.org/stable/1910129 doi: 10.2307/1910129

Lee, Y.-S., Wollack, J. A., & Douglas, J. (2009). On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educational and Psychological Measurement*, *69*(2), 181–197. doi: 10.1177/0013164408322026

Leiserson, C. C. E., Rivest, R. R. L., Stein, C., & Cormen, T. H. (2009). *Introduction to Algorithms* (3rd ed.). The MIT Press.

Levy, R. (2009). The Rise of Markov Chain Monte Carlo Estimation for Psychometric Modeling. *Journal of Probability and Statistics*, *2009*, 1–18. Retrieved from http://www.hindawi.com/journals/jps/2009/537139/ doi: 10.1155/2009/537139

Li, F., Cohen, A., Bottge, B., & Templin, J. (2016). A Latent Transition Analysis Model for Assessing Change in Cognitive Skills. *Educational and Psychological Measurement*, *76*(2), 181–204. Retrieved from http://journals.sagepub.com/doi/10.1177/0013164415588946 doi: 10.1177/0013164415588946

Liou, M., & Chang, C.-H. (1992). Constructing the exact significance level for a person fit statistic. *Psychometrika*, *57*(2), 169–181. Retrieved from http://dx.doi.org/10.1007/BF02294503 doi: 10.1007/BF02294503

Little, J. D. C., Murty, K. G., Sweeney, D. W., & Karel, C. (1963). An Algorithm for the Traveling Salesman Problem. *Operations Research*, *11*(6), 972–989. Retrieved from http://www.jstor.org/stable/167836

Liu, J., Xu, G., & Ying, Z. (2012). Data-Driven Learning of Q-Matrix. *Applied Psychological Measurement*, *36*(7), 548–564. Retrieved from http://journals.sagepub.com/doi/10.1177/0146621612456591 doi: 10.1177/0146621612456591

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli*, *19*(5),

1790–1817. Retrieved from http://arxiv.org/abs/1010.6120{\%}5Cnhttp://dx.doi.org/10.3150/12-BEJ430 doi: 10.3150/12-BEJ430

Liu, X., Han, Z., & Johnson, M. S. (2018). The UMP Exact Test and the Confidence Interval for Person Parameters in IRT Models. *Psychometrika*, *83*(1), 182–202. Retrieved from http://link.springer.com/10.1007/s11336-017-9580-y doi: 10.1007/s11336-017-9580-y

Liu, X., & Johnson, M. S. (n.d.). Estimating CDMs using MCMC. In M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models.* Springer.

Liu, X., Yang, J., Chae, H. S., & Natriello, G. (2018). Power Divergence Family of Statistics for Person Parameters in IRT Models. *Manuscript submitted for publication.*

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum Associates, Inc., 365 Broadway, Hillsdale, NJ.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*(2), 233–245. Retrieved from http://dx.doi.org/10.1007/BF02294018 doi: 10.1007/BF02294018

Lugannani, R., & Rice, S. (1980). Saddle Point Approximation for the Distribution of the Sum of Independent Random Variables. *Advances in Applied Probability*, *12*(2), 475. Retrieved from http://www.jstor.org/stable/1426607?origin=crossref doi: 10.2307/1426607

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067. Retrieved from http://doi.wiley.com/10.1002/sim.3680 doi: 10.1002/sim.3680

MacEachern, S. N. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics - Simulation and Computation*, *23*(3), 727–741. doi: 10.1080/03610919408813196

Mair, P., & Hatzinger, R. (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, *49*(1), 26–43.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal Chemical Physics.* doi: http://dx.doi.org/10.1063/1.1699114

Miyazaki, K., & Hoshino, T. (2009). a Bayesian Semiparametric Item Response Model With. *Psychometrika*, *74*(3), 375–393. Retrieved from http://link.springer.com/10.1007/s11336-008-9108-6 doi: 10.1007/s11336-008-9108-6

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*(1), 75–106. Retrieved from http://dx.doi.org/10.1007/BF02294745 doi: 10.1007/BF02294745

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122. Retrieved from https://ac.els-cdn.com/S0022249605000969/1-s2.0-S0022249605000969-main.pdf?{_}tid=0998263c-cfbb-11e7-a04e-00000aacb35d{\&}acdnat=1511378968{_}5294df6ffa5fa502afff3ee44b95f185 doi: 10.1016/j.jmp.2005.11.006

Neal, R. M. (1998). Probabilistic Inference Using Markov Chain Monte Carlo Methods. *Technical Report*, *1*, 1–144. Retrieved from papers2://publication/uuid/0C88167E-5379-4E4E-A9E4-007ABA4F716D doi: 10.1021/np100920q

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, *9*(June), 249–265. Retrieved from http://www.jstor.org/stable/1390653http://www.jstor.org/stable/1390653?seq=1{\&}cid=pdf-reference{\#}references{_}tab{_}contentshttp://about.jstor.org/terms doi: 10.1080/10618600.2000.10474879

Pan, J.-C., & Huang, G.-H. (2014). Bayesian Inferences of Latent Class Models with an Unknown Number of Classes. *Psychometrika*, *79*(4), 621–646. Retrieved from http://link.springer.com/10.1007/s11336-013-9368-7 doi: 10.1007/s11336-013-9368-7

Park, J. Y., Johnson, M. S., & Lee, Y. S. (2015). Posterior predictive model checks for cognitive diagnostic models. *International Journal of Quantitative Research in Education*, *2*(3/4), 244. Retrieved from http://www.inderscience.com/link.php?id=71738 doi: 10.1504/IJQRE.2015.071738

Plummer, M. (2005). JAGS: just another Gibbs sampler. In *Proceedings of the 3rd international workshop on distributed statistical computing (dsc 2003).*

Ritter, C., & Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy–Gibbs sampler. *Journal of the American Statistical Association.* doi: 10.1080/01621459.1992.10475289

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from http://lmr.zozlak.org/SkalowanieJednoWymiarowe/Rizopoulos{_}2006{_}ltmAnRPackageForLatentVariableModelingandIRTAnalyses.pdf

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, *7*(1), 110–120.

doi: 10.1214/aoap/1034625254

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, *12*(4), 1151–1172. Retrieved from http://projecteuclid.org/euclid.aos/1176346785 doi: 10.1214/aos/1176346785

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. Retrieved from http://projecteuclid.org/euclid.aos/1176344136 doi: 10.1214/aos/1176344136

Sethuraman, J. (1994). *A constructive definition of Dirichlet priors* (Vol. 4). Retrieved from http://www3.stat.sinica.edu.tw/statistica/j4n2/j4n27/..{\%}5Cj4n216{\%}5Cj4n216.htm doi: ***

Sijtsma, K. (1998). Methodology Review: Nonparametric IRT Approaches to the Analysis of Dichotomous Item Scores. *Applied Psychological Measurement*, *22*(1), 3–31. Retrieved from http://journals.sagepub.com/doi/10.1177/01466216980221001http://apm.sagepub.com.ugrade1.eul.edu.eg:2048/content/22/1/3{\%}0Ahttp://online.sagepub.com.ugrade1.eul.edu.eg:2048/search/results doi: 10.1177/01466216980221001

Sinharay, S. (2005). Assessing Fit of Unidimensional Item Response Theory Models Using a Bayesian Approach. *Journal of Educational Measurement*, *42*(4), 375–394. Retrieved from http://doi.wiley.com/10.1111/j.1745-3984.2005.00021.x doi: 10.1111/j.1745-3984.2005.00021.x

Skrondal, A., & Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, *34*(4), 712–745. doi: 10.1111/j.1467-9469.2007.00573.x

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc. B*, *64*(4), 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2014). *OpenBUGS User Manual* (Vol. 164). Retrieved from http://www.openbugs.net/Manuals/Manual.html

Tatsuoka, K. K. (1983). RULE SPACE: AN APPROACH FOR DEALING WITH MISCONCEPTIONS BASED ON ITEM RESPONSE THEORY. *Journal of Educational Measurement*, *20*(4), 345–354. Retrieved from http://doi.wiley.com/10.1111/j.1745-3984.1983.tb00212.x doi: 10.1111/j.1745-3984.1983.tb00212.x

Templin, J., & Bradshaw, L. (2014). Hierarchical Diagnostic Classification Models: A Family of Models for Estimating and Testing Attribute Hierarchies. *Psychometrika*, *79*(2), 317–339. Retrieved from http://link.springer.com/10.1007/s11336-013-9362-0 doi: 10.1007/s11336-013-9362-0

Templin, J., & Hoffman, L. (2013). Obtaining Diagnostic Classification Model Estimates Using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50. Retrieved from http://doi.wiley.com/10.1111/emip.12010 doi: 10.1111/emip.12010

Thissen, D. (2016). Bad Questions: An Essay Involving Item Response Theory. *Journal of Educational and Behavioral Statistics*, *41*(1), 81–89. Retrieved from http://jeb.sagepub.com/cgi/content/short/41/1/81 doi: 10.3102/1076998615621300

Thomas, A., Spiegelhalter, D. J., & Gilks, W. R. (1992). BUGS: a Program to Perform Bayesian Inference using Gibbs Sampling. In *Bayesian statistics* (Vol. 4, pp. 837–842).

van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika*, *67*(4), 519–538. Retrieved from http://link.springer.com/10.1007/BF02295129 doi: 10.1007/BF02295129

von Davier, M. (2008). A general diagnostic model applied to language testing data. *The British journal of mathematical and statistical psychology*, *61*(Pt 2), 287–307. doi: 10.1348/000711007X193957

von Davier, M. (2014). The Log-Linear Cognitive Diagnostic Model (LCDM) as a Special Case of the General Diagnostic Model (GDM). *ETS Research Report Series*. doi: 10.1002/ets2.12043

Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity. *Medical Care*. doi: 10.1097/00005650-199603000-00003

Ware, J. E. J., & Sherbourne, C. D. (1992). The MOS 36-ltem Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection. *Med Care*. doi: 10.1097/00005650-199206000-00002

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. Retrieved from http://link.springer.com/10.1007/BF02294627 doi: 10.1007/BF02294627

Wasserman, L. (2004). *All of Statistics*. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-0-387-21736-9 doi: 10.1007/978-0-387-21736-9

Xu, G., & Shang, Z. (2018). Identifying Latent Structures in Restricted Latent Class Models. *Journal of the American Statistical Association*, *113*(523), 1284–1295. Retrieved from https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1340889 doi: 10.1080/01621459.2017.1340889

Xu, G., & Zhang, S. (2016). identifiability of Diagnostic Classification Models. *Psychome-*

trika, *81*(3), 625–649. doi: 10.1007/s11336-015-9471-z

Yang, H., O'Brien, S., & Dunson, D. B. (2011). Nonparametric Bayes Stochastically Ordered Latent Class Models. *Journal of the American Statistical Association*, *106*(495), 807–817. Retrieved from `http://amstat.tandfonline.com/action/journalInformation?journalCode=uasa20http://dx.doi.org/10.1198/jasa.2011.ap10058` doi: 10.1198/jasa.2011.ap10058